Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par le

Centre international d'études supérieures en sciences agronomiques, Montpellier

Préparée au sein de l'école doctorale SIBAGHE
(Systèmes intégrés en biologie, agronomie, géosciences, hydrosciences, environnement)
et de
l'UMR AGAP, CIRAD, 34398 Montpellier, France

Spécialité : Biologie Intégrative des Plantes (BIP)

Présentée par David CROS

Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (*Elaeis guineensis* Jacq.)

Soutenue le 11/12/2014 devant le jury composé de

Mr Jacques DAVID, *Professeur, Montpellier SupAgro*	Examinateur
Mme Pascale LE ROY, *Directeur de recherches, INRA*	Examinateur
Mr Jean-Marc BOUVET, *Chercheur, CIRAD*	Directeur de thèse
Mr Leopoldo SANCHEZ, *Chargé de recherches, INRA*	Directeur de thèse
Mr Philippe BRABANT, *Professeur, AgroParisTech*	Rapporteur
Mme Zulma VITEZICA, *Maître de conférences, ENSAT*	Rapporteur
Mr Bruno Nouy, *PalmElit*	Invité

Montpellier SupAgro

Version finale après soutenance, imprimée le 3 février 2015

Remerciements

Je remercie mes directeurs de thèse, Jean-Marc Bouvet (CIRAD) et Leopoldo Sánchez (INRA Orléans) pour leurs idées, leurs conseils et leurs corrections.

Je remercie Marie Denis (CIRAD) pour toute l'aide qu'elle m'a apportée, notamment pour R, ASReml et la modélisation. Je remercie aussi Albert Flori (CIRAD) pour ses explications sur un tas de choses de génétique quantitative et Jesús Fernández (INIA, Madrid) pour sa collaboration sur le développement de la version de MOLCOANC adaptée aux plantes.

La production des données moléculaires a demandé un travail considérable, pour lequel je remercie Patrick Samper, Virginie Pomiès, Catherine Carrasco-Lacombe, Aurore Manez, Benoit Cochard et Sébastien Tisné (CIRAD). L'acquisition des données phénotypiques a représenté un travail encore plus grand. Je remercie pour cela, et pour la mise à disposition de ces données, la SOCFINDO (Indonésie), le CRAPP (Bénin) et PalmElit (Montferrier sur Lez), ainsi que Virginie Riou et Albert Flori.

Je remercie toutes les personnes qui m'ont aidées sur des points techniques délicats : Mahdi Saatchi pour la dérégression des valeurs additives, Juan Pablo Gutiérrez García et Isabel Cervantes pour les apparentements réalisés, Renaud Rincent pour l'optimisation des populations d'apprentissage par CDmean, Andres Legarra pour l'apparentement moléculaire de VanRaden avec des marqueurs multiallèliques et Eric Gozé au moment de l'installation et de l'utilisation d'ASReml-R sur le serveur linux.

Je remercie les personnes avec qui j'ai discuté à Montpellier ou à l'occasion de la présentation des études de cette thèse, et grâce à qui j'ai pu améliorer mon travail (conférences PAG [San Diego, 2013], ISOPB et IOPC [Bali, 2014], école chercheur SelGen [Rennes, 2013], comités de thèse et ateliers divers : OPGP [Montpellier, 2012 ; Kuala Lumpur, 2013 ; Bali, 2014], Novel Tree [Helsinki, 2012], ATP Sepang [Cirad, Montpellier, 2012 et 2013], etc.) : Sébastien Tisné, Albert Flori, Norbert Billotte, Jean-Marc Bouvet, Marie Denis, Cécile Grenier et Vi Cao (CIRAD), Bruno Nouy, Tristan Durand-Gasselin et Benoit Cochard (PalmElit), Léopoldo Sanchez, Hélène Muranty, Laurence Moreau, Eduardo Manfredi, Vincent Segura et Catherine Bastien (INRA), Tzer-Ying Seng (FELDA), Mukesh Sharma (Asian Agri), C.K. Wong (AAR), Jacques David et Dominique This (SupAgro), Vincent Souchard, Alexandre Marchal, etc.

Je remercie Maxime Mercière et l'INRA d'Orléans pour m'avoir prêté un bout de leur serveur pour faire mes calculs pendant quelques mois, et Alexandre Marchal pour son très bon travail lors de l'étude sur l'amélioration de la méthodologie de la sélection génomique durant son stage de Master 2.

Je remercie Laurence Dedieu (CIRAD) pour avoir relu mes deux derniers articles.

Je remercie PalmElit, et en particulier Bruno Nouy et Tristan Durand-Gasselin, pour le financement de ce travail.

Je remercie Roselyne Lannes pour son aide matérielle.

Je remercie Monique Boudet pour son aide.

Enfin, je remercie mon épouse Viviane pour son soutien.

Résumé

Etude des facteurs contrôlant l'efficacité de la sélection génomique chez le palmier à huile (*Elaeis guineensis* Jacq)

La production agricole doit augmenter à un rythme jamais atteint pour faire face à la forte hausse attendue de la demande alimentaire. La sélection génomique (SG) pourrait y contribuer en donnant la possibilité de sélectionner des individus uniquement sur leur génotype, rendant ainsi l'amélioration génétique des rendements plus efficace. L'amélioration actuelle de la production du palmier à huile, première plante oléagineuse au monde, se fait par sélection récurrente réciproque pour produire des hybrides. L'intégration de la SG à ce schéma aurait des retombées majeures. Cette thèse vise à évaluer le potentiel de la SG pour prédire les aptitudes à la combinaison hybride dans les populations parentales (Deli et groupe B).

Des données du dernier cycle d'amélioration ont permis d'obtenir la première estimation empirique de la précision de la SG. Malgré les petites populations disponibles pour calibrer le modèle génomique, cette étude a montré qu'avec des candidats à la sélection apparentés à la population de calibration (même fratrie, descendants), la précision était suffisante pour faire une présélection sur certaines composantes du rendement dans le groupe B. Par ailleurs, des simulations sur quatre générations ont montré que, pour plusieurs stratégies de SG (en particulier avec une calibration faite uniquement à la première génération en incluant des génotypes d'hybrides), la précision de sélection chez les individus non testés en croisement était suffisante pour sélectionner des parents uniquement sur leur génotype. Ceci a abouti à une augmentation de plus de 50% du gain génétique annuel par rapport à la méthode classique. Une augmentation plus rapide de la consanguinité a aussi été mise en évidence, mais elle pourrait être limitée par des méthodes classiques de gestion de la consanguinité. Finalement, les données expérimentales et simulées indiquent que la SG pourrait diminuer l'intervalle moyen de génération et accroître l'intensité de sélection, accélérant ainsi considérablement le progrès génétique sur le rendement en huile de palme.

Un schéma de sélection génomique récurrente réciproque est proposé pour le palmier à huile. Son application nécessite de confirmer expérimentalement les simulations en estimant sur plusieurs générations la précision de sélection sans recalibration du modèle. Ces futures recherches devraient utiliser les nouveaux modèles de SG, potentiellement plus efficaces (prise en compte des effets non additifs ou d'informations a priori sur les effets des marqueurs, etc.).

Mots-clés : Sélection génomique, amélioration génétique, palmier à huile, sélection récurrente réciproque, validation croisée, simulations

Abstract

Factors controlling the accuracy of genomic selection in oil palm (*Elaeis guineensis* Jacq)

Agricultural production must increase at an unprecedented rate to meet the strong growth expected in food demand. Genomic selection (GS) could contribute to reaching this goal by allowing selection of individuals on their sole genotype, making breeding more efficient. Breeding for yield in oil palm, the first oil crop in the world, is currently based on hybrid production by reciprocal recurrent selection. The integration of GS to this scheme would have major repercussions. This thesis aims to assess the potential of GS to predict hybrid combining abilities in parental populations (Deli and group B).

Data from the last breeding cycle were used to obtain the first empirical estimate of GS accuracy. Despite the small populations available to calibrate the genomic model, the study showed that with candidates related to the training population (sibs, progenies), the accuracy was sufficient to make a pre-selection in the group B on some yield components. In addition, simulations over four generations showed that the accuracy of several GS strategies (especially when training the model only in the first generation using hybrid genotypes) was high enough for non progeny tested individuals to allow selecting among them on their genotype. This resulted in an increase of more than 50% of annual genetic gain compared to traditional breeding. A faster increase in inbreeding was also demonstrated, but this could be limited by conventional methods of inbreeding management. Finally, the experimental and simulated data indicated that GS could reduce the average generation interval and increase the selection intensity, vastly speeding up the genetic progress for oil palm yield.

A recurrent reciprocal genomic selection scheme was suggested for oil palm. Its application requires an experimental confirmation of the simulations, by estimating GS accuracy over several generations without retraining the model. Future research should use new GS models, potentially more effective (taking into account non additive effects or a priori information on marker effects, etc.).

Keywords: Genomic selection, genetic improvement, oil palm, reciprocal recurrent selection, cross-validation, simulations

Sommaire

Liste des tableaux

Dans les publications (Chapitre V et Chapitre VI) :

Liste des figures

Symboles, abréviations et formules utiles

(seuls les principaux symboles et abréviations sont indiqués ici, les matrices sont en gras)

A	Valeur génétique additive
$\boldsymbol{A}$	Matrice d'apparentement additif
AGC	Aptitude générale à la combinaison
ASC	Aptitude spécifique à la combinaison
$Cov(X,Y)$	Covariance entre les variables aléatoires X et Y

$$Cov(X_1 + Y_1, X_2 + Y_2) = Cov(X_1, X_2) + Cov(X_1, Y_2) + Cov(Y_1, X_2) + Cov(Y_1, Y_2)$$

$$Cov(aX, bY) = ab\ Cov(X,Y)$$

D	Valeur génétique de dominance
$\boldsymbol{D}$	Matrice de dominance
DL	Déséquilibre de liaison
E	Effet environnemental
f_{ij}	Coefficient de parenté entre deux individus i et j
%FR	Pourcentage de fruits dans les régimes
G	Valeur génétique totale
$\boldsymbol{G}$	Matrice d'apparentement moléculaire
GEBV	*Genomic estimated breeding value* (valeur génétique additive génomique)
I	Valeur génétique d'épistasie
%HP	Pourcentage d'huile dans la pulpe
%HR	Pourcentage d'huile dans les régimes
NR	Nombre de régimes
P	Valeur phénotypique
p_b, p_B, p_{bB}	Fréquences des allèles b et B et du génotype bB, respectivement
%PF	Pourcentage de pulpe dans les fruits
PF	Poids moyen des fruits (g)

PM Poids moyen des régimes (kg)

PR Poids de régimes (kg), PR = NR × PM

φ_{ij} Coefficient de fraternité entre deux individus *i* et *j*

$r_{X,Y}$ Coefficient de corrélation linéaire de Pearson entre les variables *X* et *Y*

$r_{X,Y} = Cov(X,Y) / (\sigma_X \sigma_Y)$

σ_X Ecart-type de la variable aléatoire *X*

σ^2_X Variance de la variable aléatoire *X*

$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + Cov(X,Y)$

$\sigma^2_{aX} = a^2\sigma^2_X$

σ^2_a Variance génétique additive

σ^2_d Variance génétique de dominance

σ^2_e Variance résiduelle

σ^2_p Variance phénotypique

CHAPITRE I. INTRODUCTION GENERALE

Le monde agricole sera confronté à plusieurs défis dans la première moitié du 21ème siècle. La population mondiale devrait dépasser neuf milliards d'individus en 2050, soit une augmentation de plus de 25% par rapport à 2014. Les effets combinés de l'augmentation de la population, de l'augmentation du niveau de vie et de l'adoption de régimes alimentaires de meilleures qualités nutritionnelles dans les pays en développement devraient aboutir à une hausse de 70% de la demande alimentaire d'ici 2050. L'utilisation de certaines cultures alimentaires à des fins industrielles, notamment les biocarburants, devrait aussi se développer. Dans le même temps, les ressources naturelles, et en particulier les terres arables et l'eau, vont devenir moins disponibles. La pression sur les systèmes agricoles sera accrue par le réchauffement global, qui devrait entrainer des accidents climatiques plus fréquents (sécheresses, inondations, etc.), des températures plus élevées, des modifications dans le régime saisonnier des précipitations, etc. Enfin, même si globalement la population mondiale augmentera, cette hausse se fera uniquement dans les pays en développement, avec une évolution contradictoire entre les zones urbaines où il y aura une hausse très forte et les zones rurales où il y aura une légère baisse. En conséquence, pour répondre à la future demande alimentaire une main d'œuvre rurale moins nombreuse devra produire plus, dans des conditions moins favorables (FAO, 2009). Parmi les solutions, il est possible d'augmenter les surfaces cultivées et le rendement des surfaces déjà cultivées. Ces deux options auront des effets environnementaux, mais l'augmentation du rendement des surfaces déjà cultivées a l'avantage de beaucoup moins impacter les écosystèmes naturels.

Depuis les années 1950, les efforts fournis dans le monde sur de nombreuses cultures ont abouti à des progrès considérables en termes de rendement potentiel, de qualité nutritionnelle et de tolérance aux stress biotiques et abiotiques. Cependant, une augmentation des rendements d'ici 2050 selon le rythme observé depuis les années 1950 ne serait pas suffisante pour répondre à la future demande alimentaire (FAO, 2009). Le défi actuel est donc de faire progresser le rythme d'augmentation de la productivité agricole à un niveau jamais atteint. Pour y parvenir, il sera nécessaire de faire appel à de nouvelles technologies et à de nouvelles méthodes.

Le rendement des productions végétales a augmenté ces dernières décennies en partie grâce à l'amélioration génétique. Pour ce type de caractère complexe contrôlé par un grand nombre de gènes (caractère quantitatif), l'amélioration génétique résulte classiquement d'une sélection phénotypique. Celle-ci peut se faire sur la base de la valeur propre des individus à sélectionner (sélection massale) ou de la valeur propre d'individus qui leurs sont apparentés (sélection généalogique). Le progrès génétique d'une génération à la suivante (ΔG) peut se prédire grâce à l'équation du sélectionneur. Selon celle-ci, ΔG est égal au produit de la

précision de la sélection ($r_{\hat{A},A}$), de l'intensité de la sélection (i) et de l'écart-type génétique additif (σ_a). $r_{\hat{A},A}$ indique la fiabilité de l'estimation de la valeur génétique additive des individus, c'est à dire les effets qu'ils transmettent en moyenne à ses descendants, i la proportion d'individus sélectionnés parmi les individus évalués et σ_a la variabilité existante au sein de la population à sélectionner en termes d'effets génétiques additifs (Figure 1). Le progrès génétique s'exprime souvent annuellement, en tenant compte de l'intervalle de génération, c'est à dire du nombre d'années nécessaires pour passer d'une génération à la suivante.

Le principal problème de l'amélioration génétique des espèces végétales est que de nombreux caractères intéressants pour l'agriculteur sont fortement affectés par l'environnement (peu héritables). Ceci rend délicat l'estimation de la valeur génétique des individus à sélectionner, en particulier dans le cas de la sélection massale. Par ailleurs, des caractères ne sont pas mesurables sur certains individus, comme les caractères de production qui s'expriment uniquement chez les femelles. L'évaluation de la valeur génétique passe donc souvent par des essais spécifiques aux champs appelés tests en croisements, grâce auxquels on déduit la valeur génétique additive d'un individu à partir de la valeur propre de ses descendants. Les tests en croisements sont précis car ils permettent de contrôler les effets environnementaux, mais en général ils sont coûteux à mettre en œuvre et augmentent la durée du cycle de sélection. Par conséquent, $r_{\hat{A},A}$ est élevé mais l'intervalle de génération est grand et i est faible. Cette méthode donne de bons résultats en termes de ΔG, mais pour relever le défi actuel de nouvelles méthodes permettant d'estimer de façon plus efficace la valeur génétique d'individus candidats sont nécessaires. La sélection assistée par marqueurs (SAM) offre cette possibilité. La méthode de SAM la plus en pointe actuellement pour l'amélioration des caractères quantitatifs est la sélection génomique (Meuwissen et al., 2001) (Figure 2). Son intégration dans les programmes d'amélioration génétique des espèces végétales peut potentiellement faire augmenter le gain génétique annuel au-delà de ce qui a été atteint pendant les cinquante dernières années (Jonas et de Koning, 2013).

La SAM pour les caractères quantitatifs fait l'objet de beaucoup d'intérêt depuis plusieurs décennies. Lande et Thompson (1990) ont montré que la précision de la SAM pour des caractères faiblement héritables pouvait théoriquement dépasser de 300% la précision de la sélection phénotypique. Cette perspective très prometteuse n'a cependant pas eu les retombées espérées. En effet, cette approche initiale de la SAM reposait sur la détection de zones du génome contrôlant le caractère d'intérêt (QTL), une étape qui s'est révélée problématique pour les caractères faiblement héritables. Tout d'abord, cette approche était peu efficace pour détecter les QTL à faibles effets. En identifiant uniquement les QTL à effets forts, le sélectionneur n'avait accès qu'à une part modeste des effets génétiques contrôlant véritablement les caractères complexes. De plus, pour des raisons liées à la méthode statistique, ceci aboutissait souvent à une surestimation des effets des QTL mis en évidence. Enfin, les effets estimés se sont souvent avérés spécifiques à l'environnement ou au fond génétique (généralement étroit) de l'étude.

La sélection génomique est née dans le prolongement des approches basées sur la détection de QTL et est efficace pour les caractères quantitatifs (Meuwissen et al., 2001). Elle se base sur les progrès de la biologie moléculaire, qui ont rendu possible le génotypage d'un grand nombre d'individus avec beaucoup de marqueurs (actuellement entre plusieurs milliers

et plusieurs dizaines de milliers), et utilise de nouvelles méthodes statistiques plus efficaces pour exploiter les données phénotypiques et moléculaires. Du point de vue méthodologique, les marqueurs sont considérés comme des effets aléatoires et ils sont tous analysés simultanément dans le modèle statistique, ce qui permet de leur estimer des effets associés non biaisés et de prendre en compte les QTL ayant de petits effets. Le modèle est calibré avec un ensemble d'individus pour lesquels on dispose à la fois du génotype et d'observations (valeur propre ou estimation de la valeur génétique). L'intérêt pratique du modèle est de donner une estimation de la valeur génétique additive ou totale d'individus candidats à la sélection à partir de leur seul génotype. En fonction de la biologie de l'espèce considérée et de son schéma d'amélioration, la sélection génomique peut permettre d'accroître le progrès génétique annuel en augmentant la précision de sélection par rapport à la sélection classique, en augmentant l'intensité de sélection ou en raccourcissant l'intervalle de génération. Le potentiel de la sélection génomique est donc particulièrement élevé pour les plantes pérennes et encombrantes, chez qui l'intervalle de génération est grand (à cause d'une expression tardive des caractères d'intérêt ou du besoin de tests sur descendance préalables à la sélection) et l'intensité de sélection est souvent faible (à cause du coût élevé et de la complexité des dispositifs expérimentaux, limitant le nombre d'individus testés).

Le palmier à huile, première plante oléagineuse au monde en termes de production, fait partie des plantes pérennes jouant un rôle majeur dans l'alimentation humaine (voir l'Annexe 1 pour des illustrations sur la plante et sa filière). L'huile de palme produite à partir de la pulpe des fruits sert pour l'essentiel à l'alimentation humaine mais a aussi des débouchés non négligeables dans l'industrie (cosmétique, biocarburants, etc.). Sa production annuelle dépasse 55 Mt (USDA, 2014) et on attend, comme pour tous les produits alimentaires, une augmentation considérable de la demande. Les besoins en huile de palme devraient se situer entre 120 et 156 Mt en 2050 (Corley, 2009). Actuellement, l'amélioration génétique du rendement s'appuie sur un schéma particulier de sélection phénotypique, la sélection récurrente réciproque (SRR) (voir l'Annexe 2 pour des illustrations). Celle-ci repose sur des croisements entre des populations de palmiers à huile complémentaires pour les composantes de la production de régimes, en général la population Deli (asiatique) qui produit de gros régimes mais en nombre réduit, et une population africaine produisant des régimes moins gros mais plus nombreux. Ceci génère de l'hétérosis dans les croisements commerciaux hybrides, dont la production annuelle de régimes dépasse de plus de 25% celle des populations parentales (Gascon et de Berchoux, 1964). L'adoption dans les années 1950 de ce schéma de sélection a marqué un tournant dans l'histoire de l'amélioration génétique du palmier à huile et a permis un progrès génétique considérable, estimé à 1% par an (Durand-Gasselin et al., 2010). Cependant, l'intervalle de génération est grand (environ 20 ans) et l'intensité de sélection est faible. En effet, des tests sur descendance coûteux sont rendus indispensables par la faible héritabilité de certaines composantes du rendement et par l'impossibilité d'évaluer la valeur propre des palmiers de type pisifera (utilisés comme parents mâles des hybrides commerciaux), dont les fleurs femelles avortent.

Compte tenu de la relative facilité des méthodes actuelles de génotypage, il est envisageable de génotyper un nombre de palmiers beaucoup plus grand que le nombre que l'on parvient à tester en croisement dans un cycle de SRR conventionnelle. La stratégie d'application pratique de la sélection génomique au palmier à huile dépendra alors de sa

précision. Si la précision de la sélection génomique est modérée, on se limitera à présélectionner les individus avant de les tester en croisement selon la méthode actuelle, ce qui permettra d'éliminer les moins performants et donc d'accroître *i*. Si la précision est élevée, on pourra alors sélectionner des individus directement sur leur génotype en se passant des tests sur descendance, permettant à la fois de diminuer l'intervalle moyen de génération et d'accroître *i*. L'application de la sélection génomique chez le palmier à huile nécessite donc de comprendre les facteurs qui affectent la précision chez cette espèce, dont l'amélioration a plusieurs caractéristiques : la production d'individus hybrides entre populations génétiquement distantes, un déséquilibre de liaison fort et une faible diversité génétique dans les populations parentales, des tests en croisement lourds et un grand intervalle de génération (environ 20 ans alors que la maturité sexuelle est atteinte à 3 ou 4 ans).

La sélection génomique semble avoir le potentiel d'augmenter fortement la productivité du palmier à huile. Compte tenu de l'importance majeure de cette espèce dans l'alimentation humaine, les retombées seraient considérables pour les filières de production et de transformation, le secteur semencier et les consommateurs. Malgré l'importance des enjeux, une seule étude concernant la sélection génomique appliquée au palmier à huile a été publiée (Wong et Bernardo, 2008). Celle-ci présente des résultats prometteurs mais obtenus avec des simulations mettant en jeu des populations très différentes des populations d'amélioration existantes. Le but de cette thèse est de combler ce manque et d'évaluer en détail le potentiel de la sélection génomique par rapport à la SRR pour le rendement en huile de palme, un caractère complexe présentant de l'hétérosis mais qui est le produit de plusieurs composantes plus élémentaires, essentiellement additives. Il s'agira donc d'une étude de la sélection génomique appliquée au cas du palmier à huile, mais les conclusions devraient aussi être intéressantes pour les espèces partageant certaines caractéristiques de son amélioration, en particulier les autres plantes pérennes.

La thèse se divisera en cinq parties :

- Un rappel des concepts de base de génétique quantitative sur lesquels s'appuie la suite de la thèse (effets des gènes, variances génétiques, parenté, etc.),

- Une revue bibliographique sur les aspects directement abordés dans cette thèse, c-à-d la SRR classique appliquée actuellement au palmier à huile et la sélection génomique,

- Une étude préliminaire à la sélection génomique, destinée à préciser les caractéristiques génétiques des populations de travail : apparentement généalogique et moléculaire, taille efficace, variances génétiques, différentiation génétique entre populations, etc.
  Ce chapitre inclut une publication, placée en annexe, qui décrit l'adaptation et l'évaluation d'une méthode d'estimation de l'apparentement généalogique :

  Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M. et Fernández J., 2014. **Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population.** Theoretical and Applied Genetics, 127(4): 981‑994.

- L'estimation empirique par validation croisée de la précision de la sélection génomique et de l'effet de plusieurs facteurs qui l'affectent (méthode statistique de calcul des GEBV, caractère, population, apparentement entre populations d'apprentissage et de test).
  Ce chapitre correspond à la publication :

  Cros D., Denis M., Sánchez L., Cochard B., Flori A. et al., 2014. **Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.).** Theoretical and Applied Genetics, 1-14. DOI: 10.1007/s00122-014-2439-z

- La comparaison par simulation de la réponse à la sélection et de l'évolution des paramètres génétiques entre la SRR conventionnelle et différents scénarios de sélection génomique, sur quatre générations.
  Ce chapitre correspond à la publication :

  Cros D., Denis M., Bouvet J.-M., Sánchez L., under review. **Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm.**

CHAPITRE II. CONCEPTS DE BASE DE GENETIQUE QUANTITATIVE

Dans cette première partie, on détaillera le modèle de génétique quantitative qui relie le phénotype au génotype, en insistant sur les concepts d'effet des gènes sur les caractères, de variances génétiques, de ressemblance entre apparentés et d'hétérosis. Ensuite, on verra comment ceci permet d'estimer la valeur génétique des individus et de prédire la réponse à la sélection. Ces notions de base seront présentées dans un cadre intrapopulation et seront étendues dans le Chapitre III au contexte de l'amélioration interpopulation, qui correspond au cas du palmier à huile. Il est donc possible de lire cette thèse en commençant au Chapitre III, en se référant si besoin aux notions générales présentées ici. Le but n'est pas d'être exhaustif, car il existe un très grand nombre de modèles et de méthodes pour estimer les valeurs et les paramètres génétiques en fonction du cas de figure, mais de présenter les aspects importants utiles à la compréhension du travail effectué dans cette thèse.

Les caractères quantitatifs, tels que le rendement, la taille, etc., sont des caractères complexes dont la valeur varie de manière continue entre individus. Pour ce type de caractères, la valeur propre ou phénotype d'un individu résulte à la fois des gènes impliqués, de l'environnement et de leurs interactions. Les caractères quantitatifs sont contrôlés par un nombre important de gènes dont les effets sont variables mais globalement faibles (Hayes et Goddard, 2001; Flint et Mackay, 2009; Hill, 2010). Quelque soit l'importance d'un gène sur le phénotype, la ségrégation de ses allèles obéit aux lois Mendéliennes. Dans le cas d'un gène faible, les effets de la ségrégation de ses allèles ne peuvent pas s'observer directement et on doit alors avoir recours à des méthodes statistiques pour déduire cette ségrégation. Dans une population, la part de la variance phénotypique qui s'explique par la ségrégation des allèles d'un seul gène est généralement faible. Cependant, si on considère dans leur ensemble tous les gènes contrôlant un caractère quantitatif, la ségrégation de leurs allèles dans la population explique alors une part importante de la variance phénotypique.

La génétique quantitative est née entre la fin du 19ème siècle et le début du 20ème des travaux de F. Galton, K. Pearson, R.A. Fisher, S. Wright et J.B.S. Haldane. Elle étudie la transmission des différences individuelles pour des caractères quantitatifs et vise à estimer l'influence respective de la génétique et de l'environnement et à prédire l'évolution de la valeur des populations, en particulier sous l'effet de la sélection. Les méthodes statistiques appliquées en génétique quantitative reposent sur des modèles mathématiques qui seront présentés dans cette section, en considérant une population de taille infinie et en panmixie (reproduction au hasard) pour une espèce diploïde.

L'amélioration des plantes étudie les méthodes de création, de sélection et de fixation des caractéristiques des plantes ayant un phénotype supérieur répondant aux attentes des agriculteurs et des consommateurs.

II. A. Modèle de base de génétique quantitative

On définit les variables aléatoires suivantes : P donnant le phénotype des individus de la population, G leur valeur génétique totale et E les effets environnementaux auxquels ils sont soumis. Les variables G et E sont indépendantes et leur espérance est nulle ($E(G)=0$, $E(E)=0$).

Le phénotype P_i d'un individu i se décompose entre des effets liés à son génotype (G_i) et des effets liés à son environnement (E_i) de la manière suivante, en négligeant les interactions $G \times E$:

$$P_i = \mu + G_i + E_i \qquad [1]$$

avec μ l'espérance phénotypique de la population.

G_i correspond à l'écart entre l'espérance phénotypique des individus possédant le génotype i et l'espérance phénotypique de la population. G_i est une valeur relative, dépendante de la population étudiée.

A son tour, la valeur génétique G_i se décompose de la manière suivante :

$$G_i = A_i + D_i + I_i \qquad [2]$$

avec :

- A_i la valeur additive de l'individu i (*breeding value*). Les effets additifs sont des effets qu'un individu transmet en moyenne à ses descendants : en espérance un parent p transmet la moitié de sa valeur génétique additive A_p. En espérance sur un nombre infini d'individus, A est nulle ($E(A)=0$).
- D_i les effets de dominance issus de l'interaction entre les allèles présents à un même gène. A et D sont indépendants et $E(D)=0$.
- I_i les effets d'épistasie issus de l'interaction entre les allèles de deux gènes différents. I est indépendant de A et D et $E(I)=0$. Dans la suite de la thèse, on négligera les effets d'épistasie.

Il s'agit du modèle polygénique de base, qui n'a pas de prétention explicative au niveau des mécanismes moléculaires qui seraient en jeu. Il permet de concilier les lois de Mendel et les observations faites sur les caractères quantitatifs. Il sert en particulier à sélectionner les individus avec les plus fortes valeurs génétiques additives afin de les utiliser comme parents des variétés commerciales.

II. B. Effets des gènes sur les caractères

II. B. 1. Valeurs génotypiques d'un gène

Considérons un gène avec un allèle favorable *B* présent dans la population à une fréquence p_B et un allèle défavorable *b* présent à une fréquence p_b. Soit P_{bb} l'espérance phénotypique des individus homozygotes *bb*, P_{BB} des homozygotes *BB* et P_{Bb} des hétérozygotes, on définit les grandeurs suivantes associées au gène (Falconer et Mackay, 1996, p. 109) :

- le phénotype intermédiaire (*m*) entre les deux homozygotes : $m = 0.5(P_{BB} + P_{bb})$
- la valeur génotypique additive (*a*) qui correspond à la moitié de l'écart existant entre l'espérance phénotypique des deux homozygotes : $a = 0.5(P_{BB} - P_{bb})$
- la valeur génotypique de dominance (*d*) qui est l'écart entre le phénotype de l'hétérozygote et le phénotype intermédiaire : $d = P_{Bb} - m$

On a alors : $P_{BB} = m + a$, $P_{Bb} = m + d$ et $P_{bb} = m - a$ (voir Figure 3A).

Les paramètres *a* et *d* ont un sens biologique. Selon l'importance relative de *a* et *d*, on distingue la superdominance ($d > a$), la dominance complète ($d = a$), la dominance incomplète ($0 < d < a$) et l'additivité ($d = 0$).

II. B. 2. Effet moyen des allèles et résidus de dominance

L'espérance phénotypique de la population (μ) se calcule en tenant compte de la fréquence des différents génotypes. Dans notre exemple avec un locus biallèlique, cela donne :

$$\mu = p_{bb}P_{bb} + p_{bB}P_{bB} + p_{BB}P_{BB},$$

avec p_{bb}, p_{bB} et p_{BB} les fréquences des génotypes *bb*, *bB* et *BB*, respectivement. On peut montrer que :

$$\mu = a(p_B - p_b) + 2p_Bp_bd + m \qquad [3].$$

Chaque allèle *k* possède un effet moyen α_k (ou effet additif) correspondant à l'écart entre l'espérance phénotypique de la population et l'espérance phénotypique des individus possédant l'allèle *k* et un allèle tiré au hasard :

$$\alpha_k = \sum_{j=1}^{n} p_j(P_{kj} - \mu) \qquad [4]$$

avec *n* le nombre d'allèles existant dans la population pour le locus considéré, p_j la fréquence de l'allèle *j* et P_{kj} l'espérance phénotypique des individus de génotype *kj*.
α_k est une valeur relative dépendante de la population étudiée car elle est liée à la moyenne phénotypique et aux fréquences allèliques de la population. En particulier, un allèle favorable l'est d'autant plus que la moyenne phénotypique de la population est faible.

On note que : $\sum_{j=1}^{n} p_j\alpha_j = 0$ [5].

On définit la valeur génétique additive A_{jk} à un locus où se trouvent les allèles j et k comme la somme de l'effet moyen associé à chacun de ces deux allèles, soit :

$$A_{jk} = \alpha_j + \alpha_k.$$

Par extension, la valeur génétique additive d'un individu i se calcule sur l'ensemble des loci (gènes) contrôlant le caractère considéré :

$$A_i = \sum_{l=1}^{q} (\alpha_j + \alpha_k)_l \qquad [6]$$

avec q le nombre de gènes.

Dans notre exemple avec un locus biallèlique on peut démontrer, à partir de [4] et [3], que (Falconer et Mackay, 1996, p. 113-114) :

$$\alpha_b = -p_B(a - d(p_B - p_b)) \text{ et}$$
$$\alpha_B = p_b(a - d(p_B - p_b)).$$

L'effet moyen des allèles inclut donc une part de la valeur génotypique de dominance.
Par ailleurs, on définit l'effet de substitution des allèles d'un gène biallèlique (α) comme la différence entre l'effet moyen de chacun de ses allèles ($\alpha = \alpha_B - \alpha_b$). On peut montrer que :

$$\alpha = a - d(p_B - p_b) \qquad [7]$$

et donc que :

$$\alpha_b = -p_B\alpha \text{ et}$$
$$\alpha_B = p_b\alpha.$$

D'où :

$$A_{BB} = 2\alpha(1 - p_B)$$
$$A_{Bb} = \alpha(1 - 2p_B)$$
$$A_{bb} = -2p_B\alpha.$$

On vérifie facilement que α est la pente de la régression linéaire de A sur le nombre de copies de l'allèle favorable.

Les valeurs génétiques additives A associées aux différents génotypes observables à un locus s'obtiennent à partir de G en minimisant D^2 par la méthode des moindres carrés, faisant de D un résidu de régression linéaire, appelé résidu de dominance. Le modèle de génétique quantitative minimise donc les effets génétiques non additifs.
Dans le cas d'un locus biallèlique, les valeurs additives A associées aux génotypes bb, bB et BB s'obtiennent par une régression linéaire entre l'espérance de la valeur génétique G et le génotype (nombre de copies de l'allèle B), pondérée par la fréquence des génotypes. Les valeurs de dominance D, c'est-à-dire δ_{BB}, δ_{Bb} et δ_{BB} s'obtiennent ensuite de la manière suivante :

$$\delta_{BB} = P_{BB} - A_{BB}$$
$$\delta_{bB} = P_{bB} - A_{bB}$$
$$\delta_{bb} = P_{bb} - A_{bb}.$$

On note qu'il existe un résidu de dominance même chez les homozygotes. Par construction, les moyennes de α et δ pour un locus donné sont nulles. Les effets moyens α_B et α_b et les résidus de dominance δ_{BB}, δ_{bB} et δ_{bb} ont un sens statistique. Ils dépendent de la structure de la population.

Finalement, la valeur génétique à un locus où se trouvent les allèles *j* et *k* est :

$$G_{jk} = \alpha_j + \alpha_k + \delta_{jk}.$$

Par extension, la valeur génétique d'un individu *i* se calcule sur l'ensemble des loci contrôlant le caractère considéré par la formule :

$$G_i = A_i + D_i = \sum_{l=1}^{q}(\alpha_j + \alpha_k + \delta_{jk})_l.$$

La Figure 3B résume ces notions.

II. C. Variances génétiques

Pour un locus *l* avec *n* allèles présents dans la population, la variance génétique additive vaut :

$$\sigma_{a_l}^2 = 2\sum_{k=1}^{n} p_k \alpha_k^2$$

et la variance génétique de dominance vaut :

$$\sigma_{d_l}^2 = \sum_{i=1}^{n}\sum_{j=i}^{n} \delta_{ij}^2 p_i p_j.$$

Dans l'exemple précédent avec un gène à deux allèles on a :

$$\sigma^2_a = 2\alpha_b{}^2 p_b + 2\alpha_B{}^2 p_B = 2(-p_B\alpha)^2\, p_b + 2(p_b\alpha)^2\, p_B = 2p_b p_B \alpha^2(p_b + p_B) =$$

$$\sigma^2_a = 2p_b p_B \alpha^2$$

et on peut montrer que :

$$\sigma^2_d = \delta^2_{bb} p_b{}^2 + \delta^2_{bB} p_b p_B + \delta^2_{BB} p_B{}^2$$

$$\sigma^2_d = (2p_b p_B d)^2.$$

On note que dans le cas d'un déterminisme purement additif σ^2_d est nulle car la valeur génotypique de dominance *d* est nulle. A l'inverse, même pour une dominance complète σ^2_a n'est pas nulle, car elle dépend de la valeur génotypique de dominance *d*, qui intervient dans le calcul de l'effet de substitution des allèles α. C'est uniquement dans le cas d'un déterminisme génétique impliquant de la superdominance que la variance de dominance σ^2_d peut représenter une part prépondérante de la variance génétique totale σ^2_g (Verrier et al., 2001, p. 64-65).

Sous l'hypothèse d'équilibre panmictique à chaque locus et d'équilibre de liaison entre loci, la variance additive sur l'ensemble des *q* loci est la somme des variances additives à chaque locus :

$$\sigma_a^2 = 2\sum_{l=1}^{q}\sum_{k=1}^{n}(p_k \alpha_k^2)_l \qquad [8].$$

Sous l'hypothèse d'équilibre de liaison entre loci, la variance de dominance sur l'ensemble des *q* loci est la somme des variances de dominance à chaque locus :

$$\sigma_d^2 = \sum_{l=1}^{q}\left(\sum_{i=1}^{n}\sum_{j=i}^{n} \delta_{ij}^2 p_i p_j\right)_l \qquad [9].$$

D'après [1] et [2] et compte tenu de l'indépendance entre *A*, *D* et *E*, la variance phénotypique vaut $\sigma^2_p = \sigma^2_g + \sigma^2_e$ et $\sigma^2_g = \sigma^2_a + \sigma^2_d$.

II. D. Ressemblance entre apparentés

Le degré de ressemblance phénotypique entre deux individus dépend de leur niveau de parenté, d'éventuels facteurs environnementaux communs aux deux individus et du caractère considéré. Deux individus apparentés se ressemblent plus que deux individus choisis au hasard dans la population. Le concept de parenté est utilisé pour définir les covariances génétiques entre individus, permettant d'estimer leurs valeurs génétiques à partir de leurs valeurs phénotypiques.

Le calcul de la parenté fait intervenir les probabilités d'identité des allèles. Lorsqu'à un locus donné deux allèles sont identiques entre deux individus on parle d'identité par état (*identity by state*, IBS). Dans le cas où ces deux allèles sont identiques car ils sont les copies d'un allèle originel transmis depuis un ancêtre récent commun aux deux individus, on parle d'identité par descendance (*identity by descent*, IBD). Deux individus sont apparentés s'ils possèdent des allèles IBD. Le coefficient de parenté f_{ij} entre deux individus i et j (*coancestry* ou *kinship*) (Wright, 1922; Malécot, 1948) indique la probabilité d'obtenir deux allèles IBD lors du tirage aléatoire d'un allèle à un même locus chez i et j. Il mesure la proportion attendue d'allèles IBD entre i et j et est compris dans l'intervalle [0,1]. Une autre manière de considérer le problème est d'imaginer un descendant k de i et j. Son coefficient de consanguinité F_k (*inbreeding coefficient*) est la probabilité d'avoir deux allèles IBD à un locus donné de k. F_k est la proportion d'allèles IBD chez k et $F_k = f_{ij}$. A titre d'exemple, f_{ij} est égal à 0 entre deux individus n'ayant aucun ancêtre commun, à 0.25 entre un individu et un de ses parents (non apparenté au second parent et non consanguin) ou entre deux pleins-frères issus de parents non apparentés, à 0.5 entre deux individus d'un même clone non consanguin et à 1 entre deux individus de la même lignée pure (Tableau 1). Par ailleurs, l'auto-parenté f_{kk} d'un individu k vaut $0.5(1 + F_k)$ et varie donc de 0.5 à 1 selon son degré de consanguinité. Les coefficients de parenté et de consanguinité font implicitement références à une population de base dans laquelle il n'y a pas d'IBD entre les allèles, c'est-à-dire une population de base composée d'individus qui sont non apparentés entre eux et non consanguins. Les coefficients de parenté et de consanguinité sont des valeurs relatives par rapport à cette population de base.

La méthode traditionnelle de calcul du coefficient de parenté se base sur le pédigrée. Il s'agit d'une parenté attendue, qui est vraie en moyenne sur un grand nombre de paires d'individus de même degré de parenté. La population de base à laquelle les résultats font référence est composée des fondateurs du pédigrée, c'est-à-dire des individus pour lesquels aucun parent n'est connu, et qui sont alors supposés non apparentés entre eux et non consanguins. Les méthodes basées sur le pédigrée supposent un enregistrement correct de la généalogie. La parenté généalogique se calcule généralement à partir de la méthode des chaînes de parenté (Wright, 1922) ou de la méthode tabulaire (Emik et Terrill, 1949). Cette dernière a l'avantage d'être simple à mettre en œuvre, même pour un grand nombre d'individus. Pour deux individus i (de parents M_i et P_i) et j (de parents M_j et P_j), il suffit d'utiliser de manière récursive les formules $f_{ii} = 0.5(1 + F_i)$ [vue précédemment] et

$f_{ij} = 0.5(f_{iMj} + f_{iPj})$. En considérant toutes les paires d'individus depuis les individus les plus anciens aux plus récents, on obtient facilement l'ensemble des coefficients de parenté.

Le coefficient d'apparentement (*relationship* ou *relatedness*) est égal au double du coefficient de parenté. Il est utilisé pour produire la matrice d'apparentement additif ***A*** (*additive relationship matrix* ou *numerator relationship matrix*). Il s'agit d'une matrice symétrique carrée dont la diagonale contient les coefficients d'auto-apparentement ($2f_{ii}$) et les éléments hors-diagonale correspondent aux coefficients d'apparentements ($2f_{ij}$).

Le coefficient de fraternité φ_{ij} indique la probabilité qu'à un locus donné les deux allèles que possède l'individu *i* soient IBD aux deux allèles que possède l'individu *j*. On peut obtenir une approximation de φ_{ij} à partir des f_{ij} par la formule (Lynch et Walsh, 1998, p. 140) :

$$\varphi_{ij} \approx f_{Mi\,Mj} f_{Pi\,Pj} + f_{Mi\,Pj} f_{Pi\,Mj} \qquad [10]$$

où M_i, P_i, M_j et P_j sont respectivement la mère et le père des individus *i* et *j*. La matrice de dominance ***D*** (*dominance relationship matrix*) est définie à partir des φ_{ij} : $\boldsymbol{D} = \{ \varphi_{ij} \}$. Les matrices ***A*** et ***D*** calculées à partir du pédigrée sont utilisées dans l'estimation des valeurs génétiques par le modèle mixte traditionnel (II. G), c-à-d ne faisant pas appel à des données moléculaires.

La parenté attendue entre deux individus peut différer de leur parenté réalisée, c'est-à-dire de la proportion effective d'allèles IBD entre eux. En effet, à la méiose se produit à chaque locus l'échantillonnage aléatoire d'un allèle parmi les deux que possède chaque parent (ségrégation Mendélienne ou *Mendelian sampling*), ce qui aura des conséquences lorsque ceux-ci ne sont pas des lignées pures. L'importance de la ségrégation Mendélienne dépend notamment du nombre de loci. Avec un seul locus, la parenté réalisée entre deux pleins-frères issus de parents non apparentés sera de 0 dans 25% des cas, 0.25 (un allèle IBD) dans 50% des cas et 0.5 (deux allèles IBD) dans 25% des cas. En moyenne la parenté est bien de 0.25 (parenté attendue), mais il existe une forte variation dans la parenté réalisée. L'importance de cette variation décroit avec le nombre de loci (voir le Tableau 2 pour une illustration).

Les données moléculaires peuvent donner accès à la parenté réalisée en tenant compte de la ségrégation mendélienne. Plusieurs méthodes de calcul ont été développées (Wang, 2014). La parenté moléculaire peut toutefois surestimer la parenté réalisée, s'il existe des allèles IBD parmi les allèles IBS et que la méthode ne permet pas de les distinguer. Avec la sélection génomique, les apparentements moléculaires sont de plus en plus utilisés dans les évaluations génétiques et ils tendent à remplacer les apparentements généalogiques. Les méthodes courantes de calcul des apparentements moléculaires utilisées dans le cadre du modèle mixte génomique (GBLUP) sont présentées à la section III. B. 3. d.

Comme *G* et *E* sont indépendantes (p. 7), la covariance phénotypique entre deux individus *i* et *j* vaut (Fisher, 1918; Malécot, 1948) :

$$Cov(P_i, P_j) = Cov(G_i, G_j) + Cov(E_i, E_j).$$

La covariance génétique est non nulle si les deux individus sont apparentés. Comme les variables *A* et *D* sont indépendants, et en l'absence d'épistasie, on a :

$$Cov(G_i, G_j) = Cov(A_i, A_j) + Cov(D_i, D_j) \qquad [11].$$

La covariance additive est non nulle si les deux individus ont des allèles IBD ($f_{ij} > 0$) et la covariance de dominance est non nulle si les deux individus possèdent des couples d'allèles IBD ($\varphi_{ij} > 0$). La covariance environnementale est non nulle si les deux individus partagent un micro-environnement commun dont l'effet ne peut être corrigé. En l'absence d'effet environnementaux communs aux individus, on a : $Cov(P_i, P_j) = Cov(G_i, G_j)$.
Enfin, en l'absence d'épistasie on peut démontrer que (Gallais, 1990, p. 122; Falconer et Mackay, 1996, p. 153; Verrier et al., 2001, p.89) :

$$Cov(G_i, G_j) = 2f_{ij}\sigma_a^2 + \varphi_{ij}\,\sigma_d^2.$$

L'amélioration génétique du palmier à huile repose sur des hybrides entre deux populations non apparentées. Le calcul de la covariance génétique dans ce cas particulier est un peu plus compliqué et sera traité dans la partie III. A. 2.

II. E. Corrélations entre caractères

Des corrélations (positives ou négatives) peuvent exister entre certains caractères. La corrélation phénotypique r_p entre deux caractères dans une population est égale au coefficient de corrélation linéaire de Pearson entre le phénotype de chacun des individus pour les deux caractères, P_1 et P_2 :

$$r_p = \frac{Cov(P_1,P_2)}{\sigma_{p_1}\,\sigma_{p_2}}$$

avec σ_{p1} et σ_{p2} l'écart-type phénotypique de chacun des caractères.

Une corrélation phénotypique peut avoir une origine génétique ou environnementale. Comme G et E sont indépendantes (p. 7), on peut décomposer la covariance phénotypique en :

$$Cov(P_1, P_2) = Cov(G_1, G_2) + Cov(E_1, E_2)$$

avec $Cov(G_1, G_2)$ la covariance génétique et $Cov(E_1, E_2)$ la covariance environnementale. De la même manière, on décompose la covariance génétique entre une covariance additive et une covariance de dominance. La corrélation additive renseigne sur la liaison entre les valeurs additives d'un individu pour les deux caractères considérés. Elle vaut :

$$r_a = \frac{Cov(A_1,A_2)}{\sigma_{a_1}\,\sigma_{a_2}}.$$

avec σ_{a1} et σ_{a2} l'écart-type additif de chacun des caractères.

Les corrélations génétiques peuvent avoir deux causes biologiques, non exclusives. Premièrement, les deux caractères peuvent être contrôlés par les mêmes gènes, qui sont alors dits pléïotropes. Une telle corrélation est difficile voire impossible à rompre. Deuxièmement, elle peut être causée par un déséquilibre de liaison entre les allèles des gènes contrôlant indépendamment les deux caractères. Ce type de corrélation peut évoluer beaucoup plus facilement avec les générations.

II. F. Hétérosis

L'hétérosis ou vigueur hybride est la supériorité phénotypique d'un croisement par rapport au phénotype du meilleur de ses parents (hétérosis meilleur parent) ou par rapport au phénotype moyen de ses parents (hétérosis parent moyen) (Gallais, 2009). L'hétérosis provient de la complémentation des génotypes parentaux. Il peut s'agir de complémentation entre des gènes différents ou de complémentation entre les allèles présents aux mêmes gènes.

La complémentation entre gènes différents correspond aux mécanismes de dominance et d'épistasie. Il y a dominance complète lorsque le génotype hétérozygote donne le même phénotype que le génotype homozygote du meilleur allèle (égalité entre valeurs génotypiques additive et de dominance, voir II. B. 1 et Figure 3). Lorsque de la dominance complète existe à plusieurs gènes, l'hybride présente forcément de l'hétérosis meilleur parent (en considérant que la dominance est toujours favorable). On peut illustrer ce mécanisme avec l'exemple de deux gènes d'allèles *A* et *a* et *B* et *b*, respectivement, avec *A* et *B* les allèles favorables. Supposons que les génotypes *AA* et *BB* ont chacun une valeur phénotypique de 4, que les génotypes *aa* et *bb* ont un phénotype de 2 et que de la dominance complète existe à chaque gène. Si on croise deux parents de génotype *AAbb* et *aaBB*, alors l'hybride *AaBb* obtenu présentera de l'hétérosis meilleur parent, avec un phénotype de 8, contre 6 pour ses deux parents. Il est possible de prédire l'hétérosis chez un hybride à partir de la valeur génotypique de dominance *d* et des fréquences allèliques à chaque gène. Soit H_{F1} l'écart entre le phénotype de l'hybride et le phénotype moyen des parents et Δp la différence de fréquence de l'allèle favorable entre les deux populations parentales. En supposant l'équilibre de Hardy-Weinberg dans les populations parentales, alors à chaque gène k on a $H_{F1_k} = \Delta p_k^2 d_k$ et, en faisant la somme sur tous les gènes, $H_{F1} = \sum_k \Delta p_k^2 d_k$ (Gallais, 1990, p. 255; Falconer et Mackay, 1996, p. 166-167). L'hétérosis due à des effets de dominance est théoriquement fixable dans les populations parentales mais la probabilité de fixer chez un même individu l'allèle favorable à chacun des gènes tend très rapidement vers zéro à mesure que le nombre de gènes augmente. Ainsi, avec seulement dix gènes cette probabilité est inférieure à 10^{-6}. Le second mécanisme de la complémentation entre gènes différents est l'épistasie, qui correspond à l'interaction entre les allèles présents à un gène et les allèles d'un autre gène. Elle suffit à générer de l'hétérosis même en l'absence de dominance (Gallais, 2009, p. 70).

La complémentation entre des allèles aux mêmes gènes correspond à la superdominance. Il y a superdominance au niveau d'un gène lorsque le génotype hétérozygote donne un phénotype supérieur à celui du génotype homozygote du meilleur allèle (valeur génotypique de dominance supérieure à la valeur génotypique additive). Pour un caractère contrôlé par plusieurs gènes, l'hétérosis peut apparaitre sans qu'il y ait de la superdominance à chaque gène, si la somme des effets génotypiques de dominance sur l'ensemble des gènes (avec et sans superdominance) est supérieure au meilleur parent.

En matière d'hétérosis, les caractères multiplicatifs, c-à-d résultant du produit entre des composantes de base, représentent un cas particulier. Chez les plantes, on peut citer comme caractère multiplicatif la production qui est le produit du nombre de fruits par leur poids moyen et la croissance qui est le produit du nombre de nœuds par la longueur des entre-nœuds. L'hétérosis chez les caractères multiplicatifs peut s'expliquer par un modèle sans dominance, par le produit de composantes purement additives et complémentaires entre les

parents (Schnell et Cockerham, 1992; Gallais, 2009, p. 68-71). Un tel modèle décrit par exemple assez fidèlement l'hétérosis du palmier à huile pour la production de régimes. Gascon et de Berchoux (1964) et Gascon et al. (1966) ont comparé des palmiers à huile des populations Deli et La Mé à leur hybride. Ils ont conclu que le nombre et le poids moyen des régimes étaient essentiellement additifs mais qu'il existait un important effet d'hétérosis pour la production de régimes qui pouvait se prédire par un modèle multiplicatif, la production de régimes des hybrides étant approximativement égale au produit de la moyenne des parents pour le poids moyen et le nombre de régimes (voir III. A. 3. d). Enfin, de l'hétérosis, même faible, chez les composantes de base du caractère multiplicatif peut entrainer une hétérosis importante chez ce dernier (Schnell et Cockerham, 1992).

II. G. Le modèle mixte appliqué à l'évaluation génétique

Le modèle mixte est un modèle statistique reliant des observations à des effets fixes et à des effets aléatoires. Henderson (1950) a développé une méthode d'analyse des modèles mixtes donnant les solutions des effets fixes (BLUE, pour *best linear unbiased estimators*) et aléatoires (BLUP, pour *best linear unbiased predictors*).

Dans les évaluations génétiques, le modèle mixte est utilisé pour prédire un vecteur de valeurs génétiques aléatoires non observables (u) à partir d'un vecteur de données (y). Les valeurs génétiques sont généralement la valeur génétique additive des individus observés ou l'aptitude générale à la combinaison de leurs parents (définie au III. A. 2. a) (Piepho et al., 2008). Le modèle mixte linéaire peut s'écrire :

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}u + e$$

avec y ($n \times 1$) le vecteur des observations, β ($p \times 1$) le vecteur des effets fixes et $\boldsymbol{X}$ ($n \times p$) sa matrice d'incidence, u ($q \times 1$) le vecteur des effets aléatoires et $\boldsymbol{Z}$ ($n \times q$) sa matrice d'incidence, et e ($n \times 1$) le vecteur des erreurs résiduelles, avec n le nombre d'observations et p et q le nombre d'effets fixes et aléatoires à estimer, respectivement.
Il suppose que u et e suivent des lois normales indépendantes :

$$\begin{bmatrix} u \\ e \end{bmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{V_u} & 0 \\ 0 & \boldsymbol{V_e} \end{bmatrix}\right)$$

avec $\boldsymbol{V_u}$ et $\boldsymbol{V_e}$ les matrices de variance-covariance génétique et résiduelle, respectivement.
Ceci implique que $y \sim N(\boldsymbol{X}\beta, \boldsymbol{V_p})$, avec une matrice de variance-covariance phénotypique $\boldsymbol{V_p} = \boldsymbol{Z}\boldsymbol{V_u}\boldsymbol{Z}^{\boldsymbol{T}} + \boldsymbol{V_e}$. Eventuellement, les données doivent donc être transformées de manière à suivre une loi normale. La matrice de variance-covariance associée à u est traditionnellement $\boldsymbol{V_u} = \boldsymbol{A}\sigma^2_a$, avec $\boldsymbol{A}$ la matrice d'apparentement généalogique impliquant les individus observés et leurs ascendants présents dans le pédigrée (II. D), et σ^2_a la variance génétique additive. Plus récemment, la méthode a été étendue pour utiliser des données moléculaires (approche génomique) avec $\boldsymbol{V_u} = \boldsymbol{G}\sigma^2_a$, où $\boldsymbol{G}$ est la matrice d'apparentement génomique calculée à partir du génotype des individus observés (II. D). Dans cette partie on ne considère que l'approche BLUP traditionnelle. L'approche génomique sera traitée plus loin (III. B. 3. d). La matrice de variance-covariance résiduelle ($n \times n$) est $\boldsymbol{V_e} = \boldsymbol{I}\sigma^2_e$ avec $\boldsymbol{I}$ matrice identité. Les composantes de la variance ($\boldsymbol{V_u}$ et $\boldsymbol{V_e}$) doivent être estimées avant de pouvoir obtenir les solutions pour les

effets fixes (β) et aléatoires (u). Par convention, les solutions des effets fixes sont nommées estimateurs (BLUE, $\hat{\beta}$) et les solutions des effets aléatoires prédicteurs (BLUP, $\hat{u}$). La méthode du maximum de vraisemblance restreint (REML) estime les variances, qui renvoient à la population de base dont descendent les individus observés. Elles sont ensuite utilisées pour prédire les effets aléatoires et estimer les effets fixes, grâce aux équations du modèle mixte d'Henderson (1984, 1986) :

$$\begin{bmatrix} \boldsymbol{X^T V_e^{-1} X} & \boldsymbol{X^T V_e^{-1} Z} \\ \boldsymbol{Z^T V_e^{-1} X} & \boldsymbol{Z^T V_e^{-1} Z + V_u^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T V_e^{-1}} y \\ \boldsymbol{Z^T V_e^{-1}} y \end{bmatrix}.$$

Comme $\boldsymbol{V_e^{-1}}$ est une matrice identité, elle peut se mettre en facteur pour donner une forme simplifiée (Mrode, 2005, p. 41) qui s'écrit, pour une analyse basée sur le pédigrée :

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T Z} \\ \boldsymbol{Z^T X} & \boldsymbol{Z^T Z + A^{-1}} \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T} y \\ \boldsymbol{Z^T} y \end{bmatrix} \qquad [12],$$

avec $\lambda = \sigma^2_e / \sigma^2_a$.

Enfin, les BLUE des effets fixes sont : $\hat{\beta} = (\boldsymbol{X^T \hat{V}_p^{-1} X})^{-1}\, \boldsymbol{X^T \hat{V}_p^{-1}} y$ et les BLUP des effets aléatoires sont : $\hat{u} = \boldsymbol{\hat{V}_u Z^T \hat{V}_p^{-1}}\, (y - \boldsymbol{X\hat{\beta}})$.

Le nom du BLUP provient de ses caractéristiques :

- il maximise la corrélation entre les valeurs additives vraies (u) et prédites ($\hat{u}$) [meilleur]. Autrement dit, il minimise $\sigma^2_{(u - \hat{u})}$, la variance des erreurs de prédictions (PEV, pour *prediction error variance*),
- les solutions $\hat{\beta}$ et $\hat{u}$ sont des fonctions linéaires des observations y [linéaire],
- $E(u - \hat{u}) = 0$ [non biaisé].

La méthode du BLUP appliquée aux évaluations génétiques présente plusieurs avantages qui l'ont rendu très populaire :

- elle utilise la covariance additive entre tous les individus observés, quelque soit leur niveau d'apparentement (dans la mesure où il est non nul), pour améliorer la précision de l'estimation des composantes de la variance et des valeurs additives. Cette matrice rend aussi compte de l'histoire de la population : sélection, contribution inégale des individus du pédigrée, etc.,
- elle gère facilement les effets fixes,
- elle gère facilement les dispositifs expérimentaux déséquilibrés,
- elle est flexible et peut tenir compte de nombreux effets : corrélations génétiques et résiduelles entre caractères, mesures répétées, groupes génétiques, variances hétérogènes, effets maternels, interactions entre génotypes et environnement, etc.

Cette méthode possède cependant certains inconvénients (Piepho et al., 2008) :

- elle suppose que les composantes de la variance soient connues sans erreur. Dans la pratique, on les estime souvent grâce au REML avant de les utiliser pour obtenir les solutions $\hat{\beta}$ et $\hat{u}$. Par conséquent, l'erreur liée à l'estimation des variances n'est pas prise en compte et elle introduit une erreur dans les solutions (notons qu'une approche Bayésienne peut régler ce problème).
- les BLUP des individus sont reserrés autour de la moyenne de leurs parents, ce qui augmente la probabilité de sélectionner des individus apparentés et abouti à un

accroissement de la consanguinité plus fort qu'avec une sélection ne tenant pas compte des apparentements,

- elle repose sur plusieurs hypothèses fréquemment non vérifiées. Notamment, elle suppose que le pédigrée renvoie à une population de base idéale, composée d'individus non apparentés et non sélectionnés, et qu'il reflète donc toute l'histoire de la population (Piepho et al., 2008). Dans la pratique, la méthode considère comme population de base les ascendants les plus lointains jusqu'où remonte le pédigrée (c-à-d sans parents connus). Des erreurs dans le pédigrée peuvent aussi biaiser les résultats. Par ailleurs, le BLUP suppose que les termes de ségrégation mendélienne (II. D) soient aléatoires, ce qui signifie notamment qu'il ne doit pas y avoir eu de présélection des individus. Cependant, le BLUP a montré une grande robustesse face à la violation de ces hypothèses, ce qui a contribué à son succès.

II. H. Héritabilité et précision de sélection

L'héritabilité au sens strict h^2 d'un caractère est la fraction de la variation phénotypique entre les individus d'une population qui est due à la variation entre leurs valeurs additives :

$$h^2 = \sigma^2_a / \sigma^2_p \qquad [13].$$

h^2 est aussi la pente de la régression entre le phénotype moyen des parents et le phénotype moyen de leurs descendants. Il ne s'agit pas d'une valeur absolue, car elle dépend de la population et de l'environnement au sens large, incluant aussi bien des effets météorologiques, pédologiques, etc. que le dispositif expérimental et le protocole d'observations. En réalité, h^2 est un cas particulier lié au concept plus général de précision de sélection.

La précision de sélection $r_{A,\hat{A}}$ (*accuracy of selection*) est la corrélation entre la valeur additive, vraie et inconnue, et son estimateur. Elle traduit la qualité de l'estimateur avec lequel sera faite la sélection. L'estimateur peut être le phénotype, un phénotype moyen calculé sur des mesures répétées, un BLUP, etc. La précision de sélection sert à prédire la réponse à la sélection (voir II. I) et peut être utilisée pour comparer des stratégies d'amélioration génétique. La fiabilité (*reliability*), terme que l'on rencontre en particulier dans l'amélioration animale, est le carré de la précision de sélection. Dans le cas le plus simple, l'estimateur de la valeur additive (A) d'un individu est son phénotype (P) représenté par une seule observation. La précision de sélection est alors la corrélation entre la valeur additive et le phénotype, soit :

$$r_{A,P} = \frac{Cov(A,P)}{\sigma_a \sigma_p} = \frac{Cov(A,\ A+D+E)}{\sigma_a \sigma_p} = \frac{\sigma^2_a}{\sigma_a \sigma_p} = \frac{\sigma_a}{\sigma_p} = h.$$

La racine de l'héritabilité au sens strict est donc la précision d'une sélection massale faite sur la base d'une unique observation par individu et h^2 est la fiabilité de ce type de sélection. Si on considère un modèle mixte appliqué à l'évaluation génétique dans lequel $\hat{u}$ est le BLUP de la valeur additive u des individus, la précision du BLUP $r_{u,\hat{u}}$ est la précision de la sélection. La précision du BLUP s'obtient à partir de la variance des erreurs de prédiction de u (PEV, pour *prediction error variance*) et de la diagonale de la matrice de variance-covariance $\boldsymbol{V_u}$ associée à u dans le modèle mixte (Clark et al., 2012, p. 8). Le BLUP $\hat{u}$ de u est tel que, pour un

individu i, $u_i = \hat{u}_i + \varepsilon_i$, avec $\hat{u}_i$ indépendants de ε_i et $\sigma^2_{\varepsilon i} = PEV$. La PEV correspond donc à la variance de la distribution (conceptuelle) des erreurs $\varepsilon_i = u_i - \hat{u}_i$ associées au BLUP des u_i. Compte tenu de l'indépendance de $\hat{u}_i$ et de ε_i, $\sigma^2_{ui} = \sigma^2_{\hat{u}i} + PEV_{ui}$. Autrement dit, la PEV représente la fraction de la variance additive qui n'a pas été prise en compte par le modèle. Par ailleurs, comme par construction $\sigma^2_{u_i} = V_{u_{ii}}$, on obtient :

$$r_{u_i,\hat{u}_i} = \sqrt{\frac{Cov(u_i,\hat{u}_i)^2}{\sigma^2_{u_i}\sigma^2_{\hat{u}_i}}} = \sqrt{\frac{Cov(\hat{u}_i+\epsilon_i,\hat{u}_i)^2}{\sigma^2_{u_i}\sigma^2_{\hat{u}_i}}} = \sqrt{\frac{\sigma^2_{\hat{u}_i}}{\sigma^2_{u_i}}} = \sqrt{\frac{\sigma^2_{u_i} - PEV_{u_i}}{\sigma^2_{u_i}}}$$

soit :

$$r_{u_i,\hat{u}_i} = \sqrt{1 - \frac{PEV_{u_i}}{V_{u_{ii}}}} \qquad [14].$$

Les PEV s'obtiennent facilement du modèle mixte (Walsh, 2013). Considérons que l'inverse de la partie la plus à gauche des équations du modèle mixte d'Henderson (éq. [12]) soit :

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T Z} \\ \boldsymbol{Z^T X} & \boldsymbol{Z^T Z + G^{-1}} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{C_{11}} & \boldsymbol{C_{12}} \\ \boldsymbol{C_{12}^T} & \boldsymbol{C_{22}} \end{bmatrix}.$$

Dans ce cas, la diagonale de $\boldsymbol{C_{22}}$ est composée de la PEV de chaque BLUP. Certains logiciels (ex : R-ASReml) renvoient l'erreur type (*standard error*) des BLUP, qui correspond à la racine de la PEV.

On peut appliquer facilement la formule [14] à toutes les évaluations génétiques faites par le modèle mixte. Par exemple, les valeurs additives prédites par un modèle individuel basé sur le pédigrée sont associées à une matrice de variance-covariance additive $\boldsymbol{A}\sigma^2_a$ dont la diagonale vaut $(1+F_i)\sigma^2_a$ (cf II. D). Dans ce cas, la précision de sélection pour un individu i est (Gilmour et al., 2009, p. 227; Daetwyler et al., 2013, p. 364) :

$$r_{u_i,\hat{u}_i} = \sqrt{1 - \frac{PEV_{u_i}}{(1+F_i)\,\sigma^2_a}} \qquad [15].$$

avec PEV_{ai} la variance des erreurs de prédiction de la valeur additive de i, σ^2_a la variance génétique additive et F_i le coefficient de consanguinité.
Enfin, on note que la supériorité de la fiabilité de la sélection $r^2_{u,\hat{u}}$ par rapport à h^2 traduit l'effet de la prise en compte des apparentements (Walsh, 2013, p. 8).

L'application du modèle mixte à des données phénotypiques d'hybrides interpopulations pour estimer les aptitudes à la combinaison parentales, comme dans le cas du palmier à huile, est traitée dans la partie III. A. 2.

II. I. Réponse à la sélection

La réponse à la sélection ΔG (ou gain génétique) est l'écart entre le phénotype moyen des descendants des individus sélectionnés $\mu_{descendants}$ et le phénotype moyen de la population dans laquelle a été appliquée la sélection $\mu_{candidats}$ (Figure 1). On peut la prédire grâce à l'équation du sélectionneur :

$$\Delta G = \mu_{descendants} - \mu_{candidats} = i \times r_{u,\hat{u}} \times \sigma_a$$

avec :

- i l'intensité de sélection égale à S / σ_p, avec S le différentiel de sélection $S = \mu_{sélectionnés} - \mu_{candidats}$ ($\mu_{sélectionnés}$ le phénotype moyen des individus sélectionnés) et σ_p l'écart-type phénotypique,
- $r_{u,\hat{u}}$ la précision de la sélection,
- σ_a l'écart-type additif.

La réponse annuelle à la sélection est le ratio $\Delta G / L$, avec L l'intervalle de génération qui indique le nombre d'années nécessaires pour passer d'une génération à l'autre.

Dans le cas simple d'une sélection massale faite sur la base d'une observation par individu, comme $r_{A,\hat{A}} = h$ (cf II. H), la réponse à la sélection devient :

$$\Delta G = i \times r_{u,\hat{u}} \times \sigma_a = S / \sigma_p \times \sigma_a / \sigma_p \times \sigma_a = S\, h^2$$

On voit donc que la grande importance accordée à l'héritabilité pour ce genre de sélection vient du fait qu'elle est le seul paramètre génétique influant la réponse à la sélection, S étant un paramètre technique sous le contrôle du sélectionneur.

Le cas de la réponse à la sélection dans un programme d'amélioration d'hybrides est traité plus loin (III. A. 2. c).

II. J. Déséquilibre de liaison et taille efficace

Le déséquilibre de liaison (DL, *linkage disequilibrium*) et la taille efficace (N_e) sont présentés ici car ils sont des paramètres importants de la précision de la sélection génomique, conditionnant l'association entre les loci contrôlant les caractères d'intérêt et les marqueurs moléculaires.

Le DL est l'association non-aléatoire entre les allèles de différents loci au sein d'une population. Plusieurs méthodes permettent de mesurer le DL (Weir, 1979, 1996; Slatkin, 2008; Russell et Fewster, 2009). Considérons deux loci avec le premier possédant un allèle A et le second un allèle B (quelque soit le nombre d'allèles par locus). Le DL se rattache à la grandeur D_{AB}, qui est la déviation entre la fréquence observée de l'haplotype AB et sa fréquence attendue sous l'hypothèse d'indépendance des allèles A et B et de reproduction au hasard. Si la transmission des allèles est indépendante entre les deux loci, c'est-à-dire en équilibre de liaison, alors l'haplotype AB doit se rencontrer à une fréquence (p_{AB}) égale au produit de la fréquence des deux allèles impliqués ($p_A p_B$). Le déséquilibre de liaison entre les allèles A et B est l'écart entre la fréquence observée et la fréquence attendue :

$$D_{AB} = p_{AB} - p_A p_B$$

D_{AB} peut donc s'obtenir facilement à partir des fréquences haplotypiques. On note que pour des loci biallèliques $D_{AB} = -D_{Ab} = -D_{aB} = D_{ab}$.

Dans la pratique ce sont souvent des données génotypiques qui sont disponibles. Dans ce cas, on utilise généralement Δ_{AB} (delta de Burrow) au lieu de D_{AB}. Δ_{AB} peut se calculer à

partir de données génotypiques et sans faire l'hypothèse de reproduction au hasard. Dans notre exemple avec deux loci biallèliques, on a (Weir, 1979) :

$$\Delta_{AB} = \frac{n_{AB}}{n} - 2p_A p_B$$

avec n_{AB} un décompte des différents génotypes possibles (voir Russell et Fewster, 2009, p. 297), n le nombre d'individus échantillonnés et p_A et p_B la proportion d'allèles A et B dans l'échantillon. Un estimateur non biaisé de Δ_{AB} s'obtient en tenant compte de la taille n de l'échantillon : $\hat{\Delta}_{AB} = \Delta_{AB} \frac{n}{n-1}$ (Weir, 1979). Comme D_{AB} et $\hat{\Delta}_{AB}$ sont sensibles aux fréquences allèliques, il est difficile de les utiliser pour comparer le DL entre différentes paires de loci. On les transforme donc en une grandeur standardisée (r) qui correspond au coefficient de corrélation entre les allèles des loci concernés. Par ailleurs comme r peut être positif ou négatif, on lui préfère la grandeur r^2 lorsque l'on s'intéresse à l'amplitude du DL. Dans notre exemple, on a (Weir, 1996; Slatkin, 2008) :

$$r^2_{D_{AB}} = D^2_{AB} \,/\, (p_A(1-p_A)p_B(1-p_B))$$

et (Weir, 1996) :

$$r^2_{\hat{\Delta}_{AB}} = \hat{\Delta}_{AB} \,/\sqrt{(p_A(1-p_A) + (p_{AA} - p_A^2))(p_B(1-p_B) + (p_{BB} - p_B^2))}\,.$$

Plusieurs facteurs influencent le DL (Gupta et al., 2005; Mackay et Powell, 2007). Les facteurs augmentant le DL sont liés à la dérive, à la constitution de la population et à la sélection, naturelle ou artificielle : consanguinité, petite nombre d'individus (dérive génétique aléatoire), isolement reproductif entre groupes d'individus, mélange de populations (*admixture*), goulots d'étranglement (réduction extrême dans la taille de la population), etc. Le DL se réduit au fil des générations par les recombinaisons, qui rompent les haplotypes existants. Parmi les facteurs accélérant le rythme de réduction du DL, on trouve les régimes de reproduction privilégiant les individus non apparentés, un taux élevé de mutation, etc. D'autres facteurs peuvent augmenter ou diminuer le DL, selon les situations, ou peuvent augmenter le DL entre certains loci et le diminuer pour d'autres. Par exemple, les mutations peuvent aboutir à un DL élevé entre allèles mutants et à un DL faible entre allèles mutants et sauvages. Enfin, certains facteurs peuvent avoir un effet uniquement sur le DL de régions particulières du génome. Par exemple, les mutations favorables créent du DL au niveau local car les allèles mutants sélectionnés et des allèles neutres physiquement proches se répandent conjointement dans la population (autostop génétique ou *genetic hitchhiking*).

L'étendue du DL correspond à la distance physique (en paires de bases) ou génétique (en Morgans) en dessous de laquelle le DL est considéré comme significatif (par exemple supérieur à 0.1 lorsque l'on mesure le DL par le r^2 entre paires de marqueurs adjacents).

Pour les populations en équilibre de Hardy-Weinberg, les fréquences allèliques et génotypiques sont constantes, dans un état d'équilibre atteint en une seule génération, et il y a équilibre de liaison entre loci, atteint en quelques générations.

Cependant pour les populations dont la taille est finie les caractéristiques génétiques évoluent sous l'effet des changements aléatoires dans les fréquences allèliques au cours des générations (Caballero, 1994; Wang, 2005). Sur un grand nombre de loci la moyenne de ces changements est nulle, mais on peut calculer une variance des changements de fréquences allèliques. La fréquence d'un allèle à un locus atteindra finalement 1 (allèle fixé dans la population) ou 0 (allèle perdu). Cette évolution est décrite par le modèle de dérive génétique

de Fisher-Wright. Celui-ci suppose une population idéale, c-à-d de taille constante, avec des générations discrètes, se reproduisant au hasard, monoïque, de sexe ratio constant, avec une contribution égale de chaque individu à la génération suivante, sans mutation, sans migration et sans sélection. Dans ce modèle, la taille de la population (N) est le paramètre majeur conditionnant son évolution : le nombre de générations attendues avant la fixation d'un allèle à un locus est positivement corrélé avec N et la variance des changements de fréquences allèliques, le taux de réduction de l'hétérozygotie et le DL sont négativement corrélés à N. On peut utiliser ce modèle pour une population qui n'est pas idéale à condition de corriger N pour tenir compte des écarts existants entre les caractéristiques de la population réelle et les conditions idéales. Cette taille de population ajustée est la taille efficace N_e. Elle correspond à la taille d'une population idéale qui aurait donné lieu à la même dérive génétique aléatoire que la population réelle. Selon le paramètre génétique à partir duquel est mesurée la dérive génétique, il existe la N_e de variance (N_{eV}), la N_e de consanguinité (N_{eC}) et la N_e des coefficients de parenté (N_{eP}). Il s'agit d'ajustements de N visant à ce que la variance des changements de fréquences allèliques (pour N_{eV}) ou le taux d'accroissement de la consanguinité (pour N_{eC}) ou de la parenté (N_{eP}) observés dans la population réelle soient égaux à ceux d'une population idéale. A l'équilibre, N_{eV} et N_{eC} sont égaux.

Dans une population de N_e constante, l'approximation de l'espérance de la statistique r^2 de mesure du DL est $1 / (4N_e c + 1)$, avec c le taux de recombinaisons entre loci en Morgans (Sved, 1971). Le DL peut donc servir à estimer N_e (Russell et Fewster, 2009). Dans une population isolée d'une espèce monoïque avec reproduction au hasard et des générations discrètes, pour des loci indépendants ($c = 0.5$), neutres et non liés à des loci sélectionnés, le DL provient exclusivement de la dérive génétique. Dans ce cas on peut estimer N_{eC} avec la formule (Hill, 1981; Waples et Do, 2010, éq. 2) :

$$N_{eC} \approx 1 / (3 \times (E(r^2_{\hat{\Delta}_{AB}}) - 1 / n)) \qquad [16]$$

avec $r^2_{\hat{\Delta}_{AB}}$ basé sur l'estimateur non biaisé du delta de Burrow et n le nombre d'individus échantillonnés dans la population. Cette formule a fait l'objet d'une correction réduisant d'éventuels biais, par exemple pour les cas où N_{eC} ou n sont petits (Waples, 2006, p. 174). N_{eC} concerne ici la population des parents des individus échantillonnés. En utilisant des marqueurs liés et le taux de recombinaisons, le DL peut aussi être utilisé pour estimer N_{eC} de générations passées (Hayes et al., 2003).

Il est aussi possible d'estimer N_e à partir du pédigrée. Dans une population idéale, la consanguinité moyenne à une génération donnée vaut $F_t = 1 - (1 - \Delta F)^t$ et $N_e = 1 / 2\Delta F$ (Falconer et Mackay, 1996). Gutiérrez et al. (2008, 2009) ont développé le concept de N_e réalisée, en considérant pour chaque individu i d'une génération t donnée l'augmentation de consanguinité ΔF_i qui s'est opérée depuis les fondateurs du pédigrée, telle que $\Delta Fi = 1 - \sqrt[t_i]{1 - F_i}$. La N_e obtenue à partir des ΔF_i est une N_{eC} moyenne sur la période couverte par le pédigrée. L'intérêt de leur approche est de tenir compte des déviations qui ont pu exister entre les caractéristiques de la population étudiée et celles d'une population idéale, telle qu'une reproduction non aléatoire ou une contribution différente des individus d'une génération à la suivante, sous l'effet par exemple de la sélection. Cervantes et al. (2011) ont étendu cette méthode au calcul de N_{eP}.

CHAPITRE III. REVUE BIBLIOGRAPHIQUE

Dans cette partie, on présentera une méthode classique de sélection phénotypique, la sélection récurrente réciproque (SRR), puis le palmier à huile et l'application de la SRR à cette espèce. On présentera ensuite la sélection génomique et comment elle pourrait être utilisée pour accélérer le rythme du progrès génétique.

III. A. La sélection récurrente réciproque classique

III. A. 1. Principe

La sélection récurrente réciproque (SRR) (Comstock et al., 1949) est utilisée pour améliorer deux sources (populations, variétés, etc.) qui ont des caractéristiques complémentaires. Les individus de la source A sont croisés au hasard avec ceux de la source B et les croisements sont évalués dans des essais au champ. Afin de gagner du temps, cette phase d'évaluation est mise à profit pour autoféconder les individus A et B testés en croisement. L'analyse des essais au champ donne des estimations de la valeur génétique additive des parents et de la valeur de dominance des croisements évalués. Ceci permet de sélectionner les meilleurs parents qui seront utilisés pour produire les croisements commerciaux les plus performants. La génération suivante est constituée des autofécondations des parents sélectionnés ainsi que des croisements réalisés au sein de chaque source entre les parents sélectionnés (tous les croisements possibles sont faits). Les sources améliorées servent à la fois comme point de départ pour poursuivre l'amélioration grâce à un nouveau cycle et pour produire du matériel commercial. La sélection récurrente réciproque permet un gain génétique grâce aux effets additifs et de dominance. En effet, la SRR améliore la valeur additive des sources parentales en concentrant les allèles favorables au sein des deux sources et exploite l'hétérosis dans les croisements entre les sources. Celle-ci est lié à la distance génétique séparant les deux sources, qui est maintenue car elles restent indépendantes. La SRR est efficace quelque soit le niveau de dominance chez le(s) caractère(s) d'intérêt.

III. A. 2. Estimation de la valeur génétique des parents

III. A. 2. a. Modèle génétique

On considère ici la situation où l'on croise deux populations panmictiques non apparentées, selon le modèle présenté dans Stuber et Cockerham (1966) [détaillé aussi dans

de Souza Jr (1992)] et, indépendamment, par Lo et al. (1997), en négligeant les effets d'épistasie.

Même si les mêmes allèles sont présents dans les deux populations parentales, ils ont des effets différents dans la population hybride (sauf si les populations parentales sont identiques). Ainsi, pour deux populations parentales A et B, l'allèle *j* d'un gène a deux effets moyens dans la population hybride A × B, $\alpha_{j_A}^{AB}$ et $\alpha_{j_B}^{AB}$, respectivement. Ils se définissent en appliquant l'équation [4] au niveau de la population hybride et correspondent à l'effet moyen des allèles d'une population en croisement avec l'autre population. Ces effets moyens donnent à chaque individu une valeur additive en croisement avec l'autre population, ou valeur additive interpopulation, que l'on notera $A_{A_i}^{AB}$ ou $A_{B_i}^{AB}$ pour un individu *i* appartenant à la population A ou B, respectivement. On aura par exemple $A_{A_i}^{AB} = \sum_{l=1}^{q}(\alpha_{A_{i_j}}^{AB} + \alpha_{A_{i_k}}^{AB})_l$ (voir [6]). On définit les variances additives interpopulations, $\sigma_{a_A}^{2^{AB}}$ et $\sigma_{a_B}^{2^{AB}}$, en appliquant l'équation [8] avec une division par deux traduisant le fait qu'un seul allèle est transmis aux descendants hybrides (Stuber et Cockerham, 1966, appendix B) :

$$\sigma_{a_A}^{2^{AB}} = \sum_{l=1}^{q}\sum_{k=1}^{n}(p_{k_A}\alpha_{k_A}^{2^{AB}})_l \text{ et}$$
$$\sigma_{a_B}^{2^{AB}} = \sum_{l=1}^{q}\sum_{k=1}^{n}(p_{k_B}\alpha_{k_B}^{2^{AB}})_l.$$

La covariance génétique entre deux hybrides $A_1 \times B_1$ et $A_2 \times B_2$ découle de [11] et s'écrit (Stuber et Cockerham, 1966, éq. 7) :

$$Cov(G_{A_1B_1}, G_{A_2B_2}) = f_{A_1A_2}\sigma_{a_A}^{2^{AB}} + f_{B_1B_2}\sigma_{a_B}^{2^{AB}} + \varphi_{A_1B_1,A_2B_2}\sigma_d^{2^{AB}}.$$

L'utilisation des coefficients de parenté (au lieu des apparentements en intrapopulation) traduit à nouveau le fait qu'un parent ne transmet qu'un seul allèle à ses descendants hybrides (Lo et al., 1997, p. 2878-2879). Par ailleurs, d'après la formule [10] et comme les populations A et B ne sont pas apparentées, $\varphi_{A_1B_1,A_2B_2}$ se simplifie pour donner :

$$Cov(G_{A_1B_1}, G_{A_2B_2}) = f_{A_1A_2}\sigma_{a_A}^{2^{AB}} + f_{B_1B_2}\sigma_{a_B}^{2^{AB}} + f_{A_1A_2}f_{B_1B_2}\sigma_d^{2^{AB}} \qquad [17].$$

L'aptitude générale à la combinaison (AGC) d'un individu se définit comme la moitié de la valeur additive interpopulation : $AGC_{A_i} = 0.5A_{A_i}^{AB}$ et $AGC_{B_i} = 0.5A_{B_i}^{AB}$. L'AGC correspond aux effets qu'un individu transmet en moyenne à ses descendants hybrides, exprimée en écart à la valeur moyenne de la population hybride. Ainsi, dans un plan de croisements complet A × B, l'AGC d'un parent *i* est égale à l'écart entre le phénotype moyen de ses descendants hybrides et le phénotype moyen de toute la population hybride (μ_{AB}). Sous l'hypothèse de populations parentales en équilibre de Hardy-Weinberg, les variances d'AGC sont égales à $\sigma_{agc_A}^{2^{AB}} = 0.5\sigma_{a_A}^{2^{AB}}$ et $\sigma_{agc_B}^{2^{AB}} = 0.5\sigma_{a_B}^{2^{AB}}$ (Gallais, 1990, p. 337)[1]. On définit l'aptitude spécifique à la combinaison (ASC) pour un croisement $A_i \times B_i$ comme l'écart entre la valeur génétique attendue sur la base de l'AGC des parents (c-à-d $\mu_{AB} + AGC_{i_A} + AGC_{i_B}$) et la valeur génétique observée. Il s'agit d'une interaction entre les deux parents du

[1] Les variances additives interpopulations ont été définies ici comme dans Stuber et Cockerham (1966). Celles de Gallais (1990, p. 336-338) sont égales au double des nôtres, pour avoir une analogie avec les variances intrapopulations. Pour cette raison, dans Gallais (1990, p. 337) les variances d'AGC sont égales au quart des variances additives interpopulations.

croisement. La variance d'ASC est égale au quart de la variance de dominance interpopulation, c-à-d $\sigma^{2^{AB}}_{asc} = 0.25\sigma^{2^{AB}}_{d}$ avec, d'après [9], $\sigma^{2^{AB}}_{d} = \sum_{l=1}^{q}(\sum_{i=1}^{n}\sum_{j=i}^{n}\delta^{2}_{ij_{AB}}p_{i_A}p_{j_B})_{l}$ (Stuber et Cockerham, 1966, appendix B; de Souza Jr, 1992, p. 644). On note qu'en général les valeurs interpopulations de cette section diffèrent des valeurs intrapopulations du Chapitre II.

Pour un hybride entre deux individus A_1 et B_1, le modèle de base de génétique quantitative (éq. [1]) devient :

$$P_{A_1B_1} = \mu_{AB} + AGC_{A_1} + AGC_{B_1} + ASC_{A_1A_2} + E_{A_1B_1} \qquad [18]$$

avec, compte tenu de [17] :

- $AGC_{A_1} \sim N(0,\ 0.5\boldsymbol{A_A}\,\sigma^{2^{AB}}_{a_A})$ et $AGC_{B_1} \sim N(0,\ 0.5\boldsymbol{A_B}\,\sigma^{2^{AB}}_{a_B})$, avec $\boldsymbol{A_A}$ et $\boldsymbol{A_B}$ les matrices d'apparentement additif des populations parentales A et B, respectivement [2],
- $ASC_{A_1B_1} \sim N(0,\ \boldsymbol{D}\sigma^{2^{AB}}_{d})$, avec $\boldsymbol{D} = \{\ f_{A_1A_2}f_{B_1B_2}\}$ la matrice de dominance [3],
- $E_{A_1B_1} \sim N(0,\ \sigma^2_e)$.

On trouve des exemples de l'application de ce modèle à des hybrides de maïs dans Bernardo (1996) et Massman et al. (2013), et à des hybrides de palmier à huile dans Purba et al. (2001) (présenté au III. A. 3. e). C'est aussi ce modèle qui sera appliqué dans cette thèse pour estimer l'AGC des parents et l'ASC des croisements.

La variance génétique entre croisements est égale à la covariance entre plein-frères hybrides (Gallais, 1990, p. 337). D'après [17] et sous l'hypothèse de populations parentales non consanguines, elle vaut :

$$\sigma^{2^{AB}}_{g_C} = 0.5\sigma^{2^{AB}}_{a_A} + 0.5\sigma^{2^{AB}}_{a_B} + 0.25\sigma^{2^{AB}}_{d}$$

soit :

$$\sigma^{2^{AB}}_{g_C} = \sigma^{2^{AB}}_{agc_A} + \sigma^{2^{AB}}_{agc_B} + \sigma^{2^{AB}}_{asc} \qquad [19].$$

Si on met uniquement à profit la variabilité entre croisements (ce qui est le cas général chez le palmier huile), la variance génétique d'intérêt chez les hybrides est $\sigma^{2^{AB}}_{g_C}$. La part des effets de dominance dans la variance génétique observée entre croisements est alors $\sigma^{2^{AB}}_{asc}/\sigma^{2^{AB}}_{g_C}$, et ce ratio indique la pertinence d'une sélection des parents sur leur AGC.

Si la variance intracroisement est mise à profit, par exemple avec le clonage des meilleurs individus hybrides, on s'intéresse aussi à la variance génétique totale de la population hybride (inter- et intracroisement), qui vaut (Stuber et Cockerham, 1966, éq. 3) :

$$\sigma^{2^{AB}}_{g} = \sigma^{2^{AB}}_{a_A} + \sigma^{2^{AB}}_{a_B} + \sigma^{2^{AB}}_{d}.$$

[2] Ce qui est équivalent à $AGC_{A_1} \sim N(0,\ \boldsymbol{A_A}\,\sigma^{2^{AB}}_{agc_A})$ et $AGC_{B_1} \sim N(0,\ \boldsymbol{A_B}\,\sigma^{2^{AB}}_{agc_B})$. La différence porte sur les variances qui seront estimées mais cela n'affecte pas les AGC.

[3] Ce qui est équivalent à $ASC_{A_1B_1} \sim N(0,\ \boldsymbol{D}\sigma^{2^{AB}}_{asc})$, avec $\boldsymbol{D} = \{\ 4f_{A_1A_2}f_{B_1B_2}\}$.

III. A. 2. b. Précision des AGC et des ASC

Dans le modèle précédent (éq. [18]), les matrices de variance-covariance associées aux AGC sont $0.5\boldsymbol{A}\sigma_a^{2^{AB}}$, avec comme diagonale $0.5(1+F_i)\sigma_a^{2^{AB}}$. D'après [14], la précision de sélection est :

$$r_{AGC_i,\widehat{AGC}_i} = \sqrt{1 - \frac{PEV_{AGC_i}}{0.5(1+F_i)\sigma_a^{2^{AB}}}}.$$

A partir de [14], on peut adopter un raisonnement similaire pour calculer la précision des ASC. Pour un croisement *i* de parents A et B elle vaut :

$$r_{ASC_i,\widehat{ASC}_i} = \sqrt{1 - \frac{PEV_{ASC_i}}{0.25(1+F_A)(1+F_B)\sigma_a^{2^{AB}}}}.$$

Viana et al. (2011) donnent un exemple du calcul de la précision d'ASC pour des hybrides intra-populations de maïs.

III. A. 2. c. Réponse à la sélection

Dans un programme d'amélioration visant à produire des hybrides commerciaux, la réponse à la sélection ΔG après un cycle est l'écart entre le phénotype moyen des hybrides de la nouvelle génération ($P_{AB(1)}$) par rapport à ceux de la génération précédente ($P_{AB(0)}$). Pour un caractère mesuré dans des hybrides entre deux populations A et B, on peut écrire :

$$\begin{aligned}\Delta G_{AB} &= P_{AB(1)} - P_{AB(0)}\\ &= (AGC_A + AGC_B + ASC_{AB})_{(1)} - (AGC_A + AGC_B + ASC_{AB})_{(0)}\\ &= r_{AGC,\widehat{AGC}_A}\, i_A\, \sigma_{AGC_A} + r_{AGC,\widehat{AGC}_B}\, i_B\, \sigma_{AGC_B} + \Delta ASC\end{aligned}$$

La précision de l'AGC des populations parentales est donc un paramètre clé de la réponse à la sélection pour des hybrides.

Considérons maintenant la performance hybride pour un caractère multiplicatif entre deux composantes purement additives. Pour illustrer ce point, on considérera la production de régimes (PR) de palmiers à huile hybrides A × B, en considérant qu'elle est égale au produit de la moyenne des parents pour le nombre de régimes (NR) et leur poids moyen (PM) et que NR et PM sont purement additifs (ce qui est en réalité une approximation, voir II. F). Sur un cycle de sélection, on a :

$$\begin{aligned}\Delta G_{PR_{AB}} &= G_{PR_{AB(1)}} - G_{PR_{AB(0)}}\\ &= G_{NR_{AB(1)}} G_{PM_{AB(1)}} - G_{NR_{AB(0)}} G_{PM_{AB(0)}}\\ &= (G_{NR_{AB(0)}} + \Delta G_{NR_{AB}})(G_{PM_{AB(0)}} + \Delta G_{PM_{AB}}) - G_{NR_{AB(0)}} G_{PM_{AB(0)}}\\ &= G_{NR_{AB(0)}} \Delta G_{PM_{AB}} + G_{PM_{AB(0)}} \Delta G_{NR_{AB}} + \Delta G_{PM_{AB}} \Delta G_{NR_{AB}}\end{aligned}$$

Or,

$$\begin{aligned}\Delta G_{NR_{AB}} &= G_{NR_{AB(1)}} - G_{NR_{AB(0)}}\\ &= 0.5(G_{NR_{A(1)}} + G_{NR_{B(1)}}) - 0.5(G_{NR_{A(0)}} + G_{NR_{B(0)}})\end{aligned}$$

$$= 0.5(\Delta G_{NR_A} + \Delta G_{NR_B})$$

De même, $\Delta G_{PM_{AB}} = 0.5(\Delta G_{PM_A} + \Delta G_{PM_B})$

Donc,

$$\Delta G_{PR_{AB}} = 0.5 G_{NR_{AB(0)}}(\Delta G_{PM_A} + \Delta G_{PM_B}) + 0.5 G_{PM_{AB(0)}}(\Delta G_{NR_A} + \Delta G_{NR_B}) + 0.25(\Delta G_{PM_A} + \Delta G_{PM_B})(\Delta G_{NR_A} + \Delta G_{NR_B}).$$

La réponse à la sélection chez les hybrides pour PR dépend donc uniquement de la réponse à la sélection chez les populations parentales A et B pour NR et PM et de la valeur initiale des hybrides pour NR et PM.

On voit donc que même pour le caractère PR, plus complexe, la réponse à la sélection mesurée chez les hybrides dépend de la précision de l'AGC des populations parentales.

III. A. 2. d. Héritabilité

A l'issue de l'application du modèle [18], à des essais génétiques de croisements hybrides, on obtient une estimation de la variance des AGC des populations parentales. Cependant, il n'y a pas d'estimation de la variance phénotypique pour celles-ci lorsqu'elles ne sont pas évaluées au champ. On ne peut donc pas appliquer la formule standard [13]. Le calcul de h^2 pour la population hybride $\sigma^2_{a(AB)} / \sigma^2_{p(AB)}$ est possible mais peu intéressant car la sélection ne porte pas dans les hybrides évalués, qui servent uniquement à estimer l'AGC des parents. Par contre, en définissant h^2_A et h^2_B comme le carré des corrélations entre les AGC des parents A et B et le phénotype de leurs descendants hybrides, elles renseignent alors sur la capacité à prédire l'AGC d'un parent à partir du phénotype des croisements faits avec un parent de l'autre groupe. Dans ce cas, en considérant les croisements hybrides $A \times B$ on a par exemple pour la population parentale A :

$$h_A = r_{P_{AB},AGC_A} = \frac{Cov(P_{AB},AGC_A)}{\sigma_p \, \sigma_{agc_A}}$$

avec P_{AB} le phénotype d'un hybride $A \times B$, AGC_A l'AGC du parent A, σ_p l'écart type phénotypique des croisements hybrides et $\sigma_{agc(A)}$ l'écart type des AGC de la population A.

Or,

$$Cov(P_{AB}, AGC_A) = Cov(AGC_A + AGC_B + ASC_{AB} + e_{AB}, AGC_A)$$

et comme A et B ne sont pas apparentées et que les AGC et les ASC sont indépendantes,

$$Cov(P_{AB}, AGC_A) = \sigma^2_{agc(A)}$$

Donc $h_A = \sigma_{agc(A)} / \sigma_p$ et $h^2_A = \sigma^2_{agc(A)} / \sigma^2_p$,

De même pour la population B, $h^2_B = \sigma^2_{agc(B)} / \sigma^2_p$.

III. A. 3. La sélection récurrente réciproque pour le rendement chez le palmier à huile

Le rendement considéré ici est celui en huile de palme, produite à partir de la pulpe. L'Annexe 1 présente des illustrations de la plante et de la production d'huile. L'Annexe 2 présente des illustrations des activités d'amélioration génétique et de production de semences.

III. A. 3. a. La filière de l'huile de palme

La culture commerciale du palmier à huile à grande échelle a démarré en Afrique et en Asie du Sud-est au début du vingtième siècle, puis s'est développée dans toute la zone tropicale. Il s'agit aujourd'hui de la première plante oléagineuse au monde en termes de production. La production d'huile de palme à été multipliée par 3.8 entre 1990 et 2010 (Figure 4) et elle dépasse aujourd'hui 55 Mt (USDA, 2014). On s'attend à ce qu'elle continue d' augmenter très fortement car la demande devrait se situer entre 120 et 156 Mt en 2050 (Corley, 2009). L'Asie du Sud-est réalise aujourd'hui l'essentiel de la production (Figure 5A). Le palmier à huile est aussi la première plante oléagineuse pour le rendement à l'hectare. Sa surface cultivée représente seulement 7% des surfaces mondiales en oléagineux mais réalise environ 39% de la production. Le rendement moyen du palmier à huile dans le monde atteint presque 4 t d'huile par hectare et par an, soit environ 10 fois plus que le soja et quatre fois plus que le colza. Dans les environnements favorables, les plantations les plus performantes produisent plus de 6 t/ha sur plusieurs dizaines de milliers d'hectares et les meilleurs croisements évalués en essais génétiques dépassent 10 t/ha. Par ailleurs, l'huile de palme est l'huile végétale la moins couteuse à produire, avec des coûts de production inférieurs de 20% à ceux du soja, et elle peut se substituer à la plupart des autres huiles végétales (Fonds français pour l'alimentation et la santé, 2012).

L'huile de palme est utilisée à 80% dans l'alimentation humaine (huile de table, huile de friture, margarine, etc.) et à 20% dans l'industrie (savonnerie, cosmétiques, lubrifiants, etc.). Environ 1% de l'huile de palme est utilisée pour produire du biodiesel. L'huile de palme brute ou raffinée contient quasiment 100% de lipides sous forme principalement de triglycérides, constitués d'un glycérol auquel sont fixés trois acides gras. La part des acides gras saturés, acide palmitique en tête, est d'environ 50%. De ce fait, son intérêt nutritionnel fait débat (Fonds français pour l'alimentation et la santé, 2012). Cependant, la présence d'huile de palme dans un régime alimentaire équilibré ne semble pas poser de problème de santé. Les principaux consommateurs sont des pays émergents (Figure 5B). Dans certains pays, en particulier en Afrique, l'huile de palme est la principale source de corps gras dans le régime alimentaire. Elle joue alors un rôle majeur dans les apports lipidiques, énergétiques et vitaminiques. En France, la consommation moyenne d'huile de palme par habitant est relativement faible. Elle était estimée à 2 kg / personne / an en 2009, soit environ 6% de la consommation totale de lipides des adultes.

Le palmier à huile est un fort enjeu de développement pour de nombreux pays du Sud. Quand il est correctement planifié par les gouvernements et mis en œuvre par les planteurs, le développement du palmier à huile se traduit par un fort développement économique des régions concernées et par une importante réduction de la pauvreté rurale. Son exploitation repose sur des systèmes de culture très diversifiés allant de l'exploitation familiale de quelques hectares au périmètre agroindustriel de plusieurs dizaines (voire centaines) de milliers d'hectares. Plus de la moitié de l'huile de palme produite aujourd'hui provient de petites exploitations, au nombre d'environ trois millions. La culture du palmier à huile est capable de générer des revenus élevés et stables. Ainsi à Sumatra (Indonésie), le revenu moyen du travail est de 36 € / jour homme pour le palmier à huile contre seulement 1.7 € / jour homme pour le riz irrigué. En Indonésie, premier pays producteur, on estime à 25

millions le nombre de personnes vivant indirectement de l'exploitation du palmier à huile (Fonds français pour l'alimentation et la santé, 2012).

Les exigences pédo-climatiques de la culture amènent une cohabitation forcée avec des zones de très forte biodiversité : Bornéo, Bassin du Congo, Amazonie. Les plantations de palmiers sont sans conteste responsables ces dernières années de la déforestation de grandes étendues de forêts primaires ou secondaires, avec des conséquences très négatives (disparition d'habitats naturels, perte de biodiversité). Un des enjeux majeurs de la filière palmier à huile aujourd'hui est donc l'évolution vers une production durable, avec une intensification sans polluer sur les surfaces existantes, afin de limiter le besoin en surfaces des nouvelles plantations et l'impact écologique de la culture. Lorsque le développement du palmier à huile est mal géré, il risque de se traduire par la disparition de forêts à haute valeur de conservation, avec des impacts négatifs sur les populations locales et sur l'environnement. Le RSPO (*Roundtable on Sustainable Palm Oil*) est une initiative internationale multi-acteurs pour la certification et la promotion d'une huile de palme durable, mise en œuvre depuis 2008. Aujourd'hui, 1.3 Mha de plantations sont certifiés RSPO, soit 10% environ de la surface mondiale plantée.

III. A. 3. b. Caractéristiques biologiques du palmier à huile

On s'attachera ici seulement aux caractéristiques importantes pour l'amélioration génétique du rendement.

Le palmier à huile (*Elaeis guineensis* Jacquin) est une monocotylédone pérenne de la famille des Arécacées, une des plus anciennes familles de plantes à fleurs. Le genre *Elaeis* compte deux espèces, *E. guineensis* originaire d'Afrique et *E. oleifera* d'Amérique latine. Les deux espèces auraient divergé il y a environ 51 millions d'années (Singh et al., 2013). L'aire naturelle d'*E. guineensis* s'étend sur plus de 6 000 km le long de la côte Atlantique d'Afrique depuis le Sénégal jusqu'à l'Angola, et s'enfonce sur 50 à 200 km à l'intérieur des terres, et sur 2 000 km au niveau de l'équateur, dans la cuvette congolaise (Figure 6). Les deux espèces sont monoïques mais avec une reproduction rendue allogame par l'alternance des cycles mâles et femelles (dioécie temporelle). Il n'y a pas de reproduction végétative naturelle mais elle est possible, bien que délicate, par culture *in vitro*. L'espèce africaine se distingue par un rendement élevé en huile et *E. oleifera* par une forte teneur en acides gras insaturés, une croissance en hauteur lente et une résistance à certains parasites et ravageurs (pourriture du cœur, mineuse des feuilles). La quasi-totalité de la production actuelle d'huile de palme repose sur l'espèce africaine, compte tenu des rendements comparativement très faibles de l'espèce américaine. L'utilisation commerciale d'*E. oleifera* se limite à la production d'hybrides interspécifiques pour des zones où la pourriture du cœur empêche la culture de l'espèce africaine. Dans cette thèse il sera uniquement question de l'espèce *E. guineensis*.

Le palmier à huile est diploïde et possède 16 paires de chromosomes (2n=32). Son génome couvre une distance génétique d'environ 2 000 cM (Billotte et al., 2005; Seng et al., 2011; Ting et al., 2014; Ukoskit et al., 2014). Plusieurs séquences du génome nucléaire entier ont été acquises par des consortiums internationaux, dirigés par ACGT (2008), Sime Darby (2009), le MPOB ou *Malaysian Palm Oil Board* (2009) (Murphy, 2014) puis le Cirad (2014). Actuellement seule la séquence obtenue par le MPOB est accessible publiquement (Singh et

al., 2013). Elle fait état d'un génome d'une longueur approximative de 1.8 Gb, dans lequel 34 802 gènes ont été prédits par similarité avec des protéines connues. La comparaison des chromosomes a révélé l'existence de nombreuses régions dupliquées dans le génome.

Le palmier à huile est une herbe géante qui produit tout au long de l'année des feuilles, entourant le bourgeon végétatif pour former la couronne. Les feuilles mesurent 6 à 9 mètres et sont composées de plus de 300 folioles. A l'aisselle de chaque feuille se trouve une inflorescence dont le devenir dépend des conditions environnementales au cours de son développement (en particulier du bilan hydrique) et des cycles sexuels endogènes du palmier. Une inflorescence pourra avorter ou devenir mâle ou femelle. Une fois fécondées, les inflorescences femelles évoluent normalement en régimes. Un régime est constitué d'un rachis (ou pédoncule) portant des épillets, sur lesquels se trouvent les drupes (fruits à noyaux). Un régime pèse entre 5 et 50 kg et contient 500 à 4 000 drupes, selon l'âge du palmier, sa population d'origine, son environnement, etc. Un fruit pèse entre 10 et 30 g et se compose généralement d'une amande (faite d'un embryon et d'albumen), d'un endocarpe ligneux (coque), de mésocarpe (pulpe) et d'un exocarpe (peau). La pulpe des fruits fournit l'huile de palme et l'amande l'huile de palmiste, dont il ne sera pas question dans cette thèse. Chez le palmier à huile coexistent trois types, définis par la morphologie interne de leurs fruits :

- le dura : il s'agit du type prépondérant dans la nature (>90%). Ses fruits possèdent une coque épaisse (de 2.5 à 7 mm) et, par conséquent, un pourcentage de pulpe assez faible.
- le pisifera : il est très rare dans la nature (<5%). Ses fruits sont dépourvus de coque et sa pulpe renferme des fibres lignifiées qui, lors d'une coupe transversale du fruit, forment un anneau autour de l'amande. Les pisifera sont généralement improductifs car leurs régimes avortent avant maturité. Les fruits de pisifera sont donc très rares mais lorsqu'ils existent ils possèdent un pourcentage de pulpe très élevé.
- le tenera : il est très rare dans la nature (<5%). Ses fruits possèdent une coque de faible épaisseur (<2 mm), et un anneau de fibres lignifiées dans la pulpe, autour du noyau.

Le déterminisme génétique de la présence ou de l'absence de coque a été mis en évidence dans les années 1930 (Beirnaert et Vanderweyen, 1941). Ce caractère est sous le contrôle d'un gène nommé *Sh* (pour *shell*). Celui-ci possède deux allèles codominants, *Sh+* qui permet la formation d'une coque et un mutant d'effet opposé *Sh-*. Les dura sont donc de génotype *Sh+//Sh+* et les pisifera *Sh-//Sh-*. Leur hybride le tenera est hétérozygote et présente un phénotype intermédiaire.

Le rendement annuel en huile de palme d'un palmier est le produit du poids de sa production de régimes (PR) et du pourcentage d'huile dans ses régimes (%HR) (aussi nommé qualité des régimes ou taux d'extraction). Le poids total de régimes est lui-même le produit du nombre de régimes (NR) et du poids moyen des régimes (PM). Il existe une corrélation négative forte entre NR et PM (Gascon et al., 1966). Le pourcentage d'huile dans les régimes est lui aussi le produit de caractères plus simples, qui sont le pourcentage de fruits dans le régime (%FR), le pourcentage de pulpe dans les fruits (%PF) et le pourcentage d'huile dans la pulpe fraîche des fruits (%HP). Cette décomposition correspond aux caractères sur lesquels portent actuellement la sélection pour l'amélioration du rendement, mais d'autres décompositions seraient possibles, par exemple en tenant compte du pourcentage d'eau dans la pulpe et d'huile dans la pulpe sèche, ou du nombre et du poids de fruits, etc.

Les inflorescences mâles produisent 10 à 50 g de pollen. Dans les conditions naturelles celui-ci reste viable quelques jours, mais en le récoltant au moment approprié et en le stockant sous vide au congélateur, on peut facilement le conserver pendant de nombreuses années. Cette caractéristique du palmier à huile, associée aux techniques efficaces de fécondations artificielles qui ont été développées dans les années 1940, ont amené les programmes de sélection et de production de semences à s'appuyer exclusivement sur des croisements contrôlés.

La plante commence à produire lors de sa 3ème ou 4ème année, selon l'environnement, et est en général exploitée pendant une vingtaine d'années, la hauteur de l'arbre rendant alors la récolte difficile.

L'huile de palme est composée d'environ 50% d'acides gras saturés. Les principaux acides gras sont l'acide palmitique (C16:0, représentant 45% du total), l'acide oléique (C18:1, 40%) et l'acide linoléique (C18:2, 10%).

III. A. 3. c. Populations d'amélioration de palmier à huile

Bien que l'utilisation du palmier à huile par les populations d'Afrique subsaharienne soit ancestrale, cette espèce n'a pas subi une domestication marquée et il n'existe pas de types distincts « sauvage » et « cultivé ».

Le palmier à huile a été introduit en Asie du Sud-est en 1848, avec quatre plantules dura plantées dans le jardin botanique de Bogor (Java, Indonésie) à des fins ornementales (Corley et Tinker, 2003). Leur origine exacte reste inconnue mais la population asiatique qui en a découlé (Deli) a des caractéristiques génétiques et phénotypiques proches de celles des populations d'Afrique centrale.

A partir des années 1920 une sélection massale a été appliquée au sein des différentes populations (Corley et Tinker, 2003; Cochard, 2008). En Afrique, elle a été réalisée principalement par les centres de recherches coloniaux (INEAC dans l'actuelle République Démocratique du Congo [RDC], IRHO en Côte d'Ivoire et au Bénin, WAIFOR au Nigéria). En Asie elle a été faite dans les grandes sociétés de plantations d'Indonésie et de Malaisie. Ce processus s'est poursuivi jusque dans les années 1940 et a donné naissance aux populations d'amélioration modernes décrites ci-dessous, parfois nommées origines géographiques, origines génétiques ou, en anglais, BPRO pour *breeding populations of restricted origins*.

Les principales populations créées en Afrique sont La Mé (Côte d'Ivoire) et Yangambi (RDC). On trouve aussi les populations Yocoboué (Côte d'Ivoire), Sibiti (République du Congo), Ekona (Cameroun), WAIFOR (Nigéria) et Pobè (Bénin) mais elles sont beaucoup moins utilisées. La population La Mé trouve son origine dans les prospections faites dans la région de Bingerville dans les années 1920. Ceci a abouti à la sélection de 19 individus choisis car leurs fruits possédaient des proportions équilibrées entre le mésocarpe (60%), l'amande (20%) et la coque (20%). On note que ceci diffère notablement de l'idéotype moderne (Cochard, 2008). La population Yangambi est issue de plantations faites dans les années 1920 à partir de 10 à 20 tenera en pollinisation libre, incluant Djongo (« le meilleur ») du jardin botanique d'Eala et des tenera de Yawenda, Ngazi et Isangi. Une sélection a été faite sur la base du rendement en régimes puis de la qualité des régimes. Les objectifs étaient essentiellement une production de régimes élevée, un fort pourcentage de pulpe dans les

fruits, de gros fruits et une amande relativement importante. Compte tenu des exigences élevées des sélectionneurs et des performances incomparables de Djongo, la population Yangambi d'origine serait issue à plus de 70% de Djongo. La population Sibiti est fortement apparentée à la population Yangambi, dont elle dérive (Demol et al., 2002; Corley et Tinker, 2003; Cochard, 2008).

En Asie, les quatre plantules de 1848 ont donné naissance à la population Deli, dans laquelle on distingue aujourd'hui plusieurs sous populations, principalement Marihat Baris en Indonésie, SOCFIN en Indonésie et Malaisie, Serdang Avenue, Ulu Remis (ou Guthrie), Johor Labis et Elmina (dont les Dumpy) en Malaisie. Les premières activités connues de sélection de la population Deli pour le rendement en huile à partir d'observations rigoureuses datent des années 1910-1930, selon les sociétés de plantation (Corley et Tinker, 2003; Cochard, 2008). Les détails concernant cette période sont incertains (caractères sélectionnés, intensité de sélection, etc).

Par ailleurs, des échanges de matériel ont abouti à la formation de l'origine Deli Dabou (Côte d'Ivoire) à partir de graines de Deli SOCFIN et à la population AVROS (Indonésie, Malaisie) à partir de graines de Djongo.

Les populations de palmier à huile peuvent se répartir en deux groupes A et B selon les caractéristiques de production de leurs régimes (Gascon et de Berchoux, 1964). Le groupe A produit des régimes plus gros que le groupe B mais le groupe B produit un plus grand nombre de régimes. Le groupe A est composé des populations Deli et Angola, le groupe B des autres populations africaines. On peut à nouveau faire des distinctions entre populations du groupe B sur la base du phénotype, avec La Mé caractérisé par un nombre très faible de régimes et Yangambi par des régimes relativement gros. Les données moléculaires ont ensuite permis de préciser cette structure. Cochard (2008) et Cochard et al. (2009) ont étudié la diversité génétique sur un ensemble de 318 individus représentant huit pays, avec du matériel issu de prospections, de jardins botaniques et de programmes d'amélioration. Avec 14 marqueurs microsatellites (SSR), ils ont mis en évidence une structure très marquée (F_{ST} = 0.243), avec trois groupes de populations bien distincts (Figure 7). Le groupe I est constitué des origines de Côte d'Ivoire, le groupe II rassemble les origines d'Afrique Centrale, du Nigéria et du Bénin, et le Groupe III est composé des origines Deli. Ils ont aussi montré que la population Deli pouvait se diviser en deux sous-populations, une renfermant les Deli Dabou, SOCFIN et Dumpy, et l'autre les Deli Guthrie et WAIFOR.

Le déséquilibre de liaison (DL) a aussi été étudié par Cochard (2008). Il a mis en évidence que le DL était plus important dans les Deli que dans les populations africaines sur les courtes distances (inférieures à 30-35 cM). Le DL mesuré par la corrélation entre SSR (r^2) chutait à moins de 0.10 après environ 17 cM chez les Deli, 10 cM chez les La Mé et 7 cM chez les Yangambi.

III. A. 3. d. Déterminisme génétique du rendement en huile de palme

Entre 1950 et 1953 des essais ont été plantés pour évaluer des croisements intra- et interpopulations faits entre les meilleurs géniteurs de cinq plantations. Cette expérimentation, nommée « Expérience Internationale » et organisée par l'IRHO, a montré (Gascon et de Berchoux, 1964; Gascon et al., 1966; Bénard, 1965; Noiret et al., 1966) :

- qu'il existe un effet d'hétérosis dans la production de régimes : les croisements inter-groupes A × B ont une production de régimes qui dépasse de plus de 25% la production de régimes des croisements intrapopulations (Figure 8),
- que le nombre de régimes (NR) et le poids moyen des régimes (PM) sont essentiellement additifs, et que l'on peut estimer la production totale de régimes d'un croisement biparental par le produit de la moyenne des parents pour NR et PM,
- que les composantes de la qualité des régimes sont essentiellement additives,
- que la population Deli est la meilleure pour la qualité des régimes dura (pourcentage d'huile dans les régimes plus élevé),
- que la production d'huile de palme des tenera hybrides A × B dépasse de plus de 30% celle des tenera africains.

De nombreuses études ont estimé l'héritabilité au sens strict (h^2) des composantes du rendement en huile de palme en considérant la variabilité additive et phénotypique au sein des populations parentales. Il en ressort des valeurs faibles à intermédiaires selon le caractère et la population. Chez les Deli NR est plus héritable que PM ($h^2_{NR} \approx 0.5$ et $h^2_{PM} \approx 0.25$, en moyenne sur 12 valeurs en Malaisie et en Côte d'Ivoire) et PR a une h^2 faible (~0.20) (Meunier et al., 1970; Corley et Tinker, 2003, p. 169). Dans la population La Mé, PM est plus héritable que NR ($h^2_{NR} \approx 0.6$ et $h^2_{PM} \approx 0.9$, dans une étude en Côte d'Ivoire) (Meunier et al., 1970). Pour les composantes du régime, les valeurs moyennes observées sur plus de 10 analyses (Corley et Tinker, 2003, p. 175) indiquent une h^2 faible pour %FR et %HP (~0.20) et intermédiaire pour %PF, %AF et %HR (~0.40 – 0.55).

Quatre études ont mis en évidence des QTL liés au rendement en huile de palme (Rance et al., 2001; Billotte et al., 2010; Jeennor et Volkaert, 2013; Ukoskit et al., 2014). La plus aboutie est celle de Billotte et al. (2010). Ils ont utilisé une approche multi-parentale avec un dispositif expérimental composé de 375 palmiers hybrides A × B, dans un plan de croisement factoriel complet entre deux Deli et deux individus du groupe B (La Mé et Yangambi), avec un génotypage fait avec 411 SSR. Ils ont mis en évidence 76 QTL impliqués dans 24 caractères quantitatifs, dont 39 QTL impliqués dans des caractères directement liés au rendement (nombre de feuilles émises, poids moyen des régimes, pourcentage d'huile dans la pulpe, etc.).

III. A. 3. e. Mise en œuvre de la sélection récurrente réciproque classique pour le rendement chez le palmier à huile

A partir des années 1950, suite à la mise en évidence du déterminisme génétique du type de fruit (dura, tenera, pisifera), les palmiers à huile tenera ont remplacé les dura dans les plantations commerciales, amenant une augmentation de 30% du rendement en huile (Corley et Lee, 1992).

En 1957, la supériorité des hybrides A × B pour la production d'huile de palme montrée par « l'Expérience Internationale » a amené l'adoption d'un schéma de sélection récurrente réciproque (SRR) (Gascon et de Berchoux, 1964; Meunier et Gascon, 1972) inspiré du maïs (Comstock et al., 1949). Il met à profit l'hétérosis issue de la complémentarité entre groupes A et B pour les composantes de la production de régimes et permet, en utilisant des

pisifera africains pour féconder les dura Deli, de produire des croisements commerciaux de type tenera et donc de bonne qualité de régimes. Il est encore appliqué aujourd'hui.

Dans le détail (Figure 9), il s'agit de sélection récurrente réciproque (Gallais, 1990, p. 333-343, 2009, p. 235). Les candidats à la sélection sont des individus appartenant à des familles de pleins-frères au sein de chaque groupe. Leur valeur propre est mesurée pour les caractères les plus héritables (essentiellement %PF et %HP). Les meilleurs individus sont présélectionnés pour être testés en croisement avec l'autre groupe. Les croisements sont observés dans des essais, selon des dispositifs expérimentaux généralement de type blocs de Fisher complets à 5 ou 6 répétitions ou de type lattice équilibré de rang 4 ou 5. Ces essais représentent des investissements lourds : actuellement chaque individu est croisé avec 2 à 4 partenaires, chaque croisement est représenté au champ par 45 à 72 individus et les observations sont réalisées de la 3ème année (début de production) à la 10ème année. A l'issue des essais, on obtient pour chaque parent une AGC de bonne précision (environ 0.90 pour toutes les composantes du rendement dans le dernier ensemble d'essais achevé, à Aek Loba en Indonésie, selon l'estimation faite dans cette thèse au Chapitre IV). Sur la base des AGC, on effectue la sélection finale sur tous les caractères. Les individus sélectionnés produiront, par croisements au sein de chaque groupe et autofécondations, la génération suivante utilisée pour démarrer un nouveau cycle de SRR et pour produire du matériel commercial.

Deux études ont utilisé le modèle mixte pour estimer la valeur génétique des individus testés en croisements. Soh (1994) a appliqué un modèle de type parent B avec groupes génétiques pour estimer l'AGC de neuf pisifera pour NR, PM, PR, %HR et CR (vitesse de croissance en hauteur). Il a utilisé un dispositif déséquilibré de 49 croisements répartis entre trois essais avec 1 à 15 croisements par pisifera. Les précisions obtenues (formule [15]) étaient très faibles, situées entre 0.10 et 0.20. Cette faible performance est certainement liée au fort déséquilibre présent dans le dispositif, à l'absence de prise en compte d'un effet parent Deli (femelle) et à l'utilisation de variances issues de la littérature et non pas estimées au cours de l'analyse.

Purba et al. (2001) ont utilisé un modèle avec deux groupes parentaux indépendants (A et B) dans un dispositif déséquilibré comportant 401 croisements entre 154 dura (groupe A) et 135 tenera et pisifera (groupe B), répartis sur 26 essais :

$$y = \boldsymbol{X}b + \boldsymbol{Z_1}\, AGC_A + \boldsymbol{Z_2}\, AGC_B + \boldsymbol{Z_d}\, ASC_{AB} + e$$

avec y : valeur des croisements, b : effet des essais, AGC_A et AGC_B : les AGC des géniteurs des groupes A ~$N(0,\ 0.5\boldsymbol{A_A}\,\sigma^{2AB}_{a_A})$ et B ~$N(0,\ 0.5\boldsymbol{A_A}\,\sigma^{2AB}_{a_B})$, respectivement, et ASC_{AB} : les aptitudes spécifiques à la combinaison (interactions entre parents) ~$N(0,\ \boldsymbol{D}\sigma^2_d)$. Les matrices $\boldsymbol{A_A}$, $\boldsymbol{A_B}$ et $\boldsymbol{D}$ sont calculées à partir des pédigrées : $\boldsymbol{A_A} = \{\, 2f_{AiAj} \,\}$, $\boldsymbol{A_B} = \{\, 2f_{BiBj} \,\}$ et $\boldsymbol{D} = \{\, f_{AiAj} \times f_{BiBj} \,\}$, selon le modèle de Stuber et Cockerham (1966) (cf éq. [18] et [17]). $\boldsymbol{X}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ et $\boldsymbol{Z_d}$ sont des matrices d'incidence. Ils ont procédé par validation croisée pour prédire la valeur de croisements non évalués au champ pour NR, PM, PR, %HR et CR. Les corrélations obtenues étaient raisonnables, allant de 0.42 à 71.

La production à grande échelle des croisements hybrides commerciaux ne peut pas se faire directement avec les seuls individus sélectionnés, à partir desquels les quantités de semences que l'on peut obtenir sont faibles (quelques milliers de graines par an et par dura). On utilise pour cela les descendants intrapopulations des individus sélectionnés, qui sont

croisés entre eux pour reproduire les hybrides interpopulations voulus. Jacquemard et al. (1981) ont vérifié grâce à des essais plantés en Côte d'Ivoire que la reproduction d'un hybride A × B était possible en croisant des échantillons d'arbres A' et B' descendants de A et B par autofécondations. Ils ont constaté qu'il y avait une identité remarquable entre un croisement et sa reproduction, aussi bien pour le poids total de régimes et ses composantes que pour la production d'huile, et qu'il y avait une bonne concordance pour la composition des régimes. Ils ont montré qu'il n'y avait pas d'effet maternel et ont observé qu'à l'intérieur d'un hybride la variance de la reproduction n'était pas significativement différente de celle du croisement de départ. Par ailleurs, ils ont conclu que 12 dura et 12 pisifera permettaient la reproduction satisfaisante d'un croisement et que 12 à 20 croisements étaient suffisants pour constituer une reproduction dont la probabilité d'être inférieure de seulement 5% à la valeur du croisement de départ était très faible (inférieure à 5%).

Ce schéma de SRR a permis un progrès génétique considérable, estimé à 1% par an depuis son adoption (Durand-Gasselin et al., 2000). Il possède cependant plusieurs inconvénients :

- des coûts importants, en particulier de main d'œuvre,
- un temps important pour connaitre la valeur génétique des individus testés en croisement, ce qui amène à un intervalle de génération long (environ 20 ans), alors que la maturité sexuelle du palmier à huile est atteinte relativement rapidement, à 3 ou 4 ans,
- un nombre relativement réduit d'individus testés sur descendance par génération et groupe parental, inférieur à 200, qui aboutit à une faible intensité de sélection.

III. B. La sélection génomique

III. B. 1. Principe de la sélection génomique

La sélection génomique (SG) est une méthode de sélection assistée par marqueurs (SAM) efficace pour améliorer les caractères quantitatifs (Meuwissen et al., 2001). Elle se base sur un marquage dense du génome et sur des méthodes statistiques, en général le modèle mixte ou des méthodes bayésiennes, capables de valoriser simultanément l'information de tous les marqueurs pour estimer la valeur additive (GEBV, pour *genomic estimated breeding value*) des individus candidats à la sélection. La SG est aussi capable de prédire la valeur génotypique totale (additive, de dominance et épistatique) mais ce point ne sera pas traité dans cette thèse. Les marqueurs sont en général des marqueurs anonymes, c-à-d pour lesquels il n'existe pas d'information a priori indiquant qu'ils soient dans ou liés à un QTL. Ils sont tous utilisés conjointement dans les analyses, sans test de signification de leur effet. Le modèle de SG est calibré avec des individus possédant des observations (phénotype ou estimation de leur valeur additive) et un génotype. Ils constituent la population d'apprentissage ou de calibration (*training*). Les candidats à la sélection peuvent être les descendants de la population d'apprentissage ou des individus d'autres populations (Figure 10).

La principale différence entre la SG et les méthodes de SAM antérieures réside dans l'approche statistique, qui permet à la SG d'être efficace pour des caractères contrôlés par un grand nombre de gènes. En effet pour ce type de caractères, les méthodes de SAM ayant précédé la SG souffraient de plusieurs inconvénients, liés au fait que tous les marqueurs n'étaient pas considérés simultanément dans le modèle et qu'un test de signification était réalisé pour conserver ou rejeter les marqueurs. En conséquence, l'effet des QTL correctement identifiés et la variance phénotypique expliquée par ces QTL étaient surestimés, en particulier lorsque le nombre d'individus phénotypés était petit (exemple 100) (Beavis, 1994, 1998). Ceci rendait quasiment impossible la détection des QTL à faibles effets. En identifiant uniquement les QTL à effets forts, le sélectionneur n'avait accès qu'à une part modeste des effets génétiques contrôlant véritablement les caractères complexes. Pour illustrer ce problème, on peut citer le travail de Schön et al. (2004) qui, à partir d'une grande population de 976 lignées F5 issues du croisement de deux lignées de maïs, ont détecté un faible nombre de QTL, allant de 18 à 32 selon le caractère (rendement en grain, humidité des grains et hauteur). Par ailleurs, ces QTL expliquaient seulement 30% à 52% de la variance génétique de la population de test. Laurie et al. (2004) ont obtenu des résultats similaires sur la teneur en huile du maïs, à partir de 500 lignées produites à partir du croisement entre des lignées à forte et faible teneur en huile. Malgré la grande taille de la population, les 50 QTL qu'ils ont mis en évidence expliquaient seulement 50% de la variance génétique. Des résultats similaires ont aussi été observés chez les espèces forestières (Muranty et al., 2014). Ceci suggère que l'approche QTL classique n'est pas très efficace pour les caractères complexes contrôlés par un grand nombre de gènes. Par ailleurs, les effets estimés se sont souvent avérés spécifiques à l'environnement ou au fond génétique (généralement étroit) de l'étude.

Les premières études de la SG ont porté sur les bovins laitiers et elles ont montré qu'un doublement du gain génétique annuel et une réduction de 90% du coût d'évaluation des taureaux étaient envisageables (Eggen, 2012). La mise en application pratique de la SG a d'abord démarré chez les bovins laitiers de race Holstein, dont l'organisation de la filière a permis de constituer de grandes populations d'apprentissage grâce auxquelles de bonnes précisions de sélection ont été obtenues. Les premiers résultats officiels d'évaluation génomique sont parus en 2009 pour quelques populations Holstein. L'application s'est ensuite étendue aux autres races laitières. Aujourd'hui, plus de 15 pays utilisent des GEBV dans leur programme d'amélioration national et les ont homologué au niveau international (Eggen, 2012; Bouquet et Juga, 2013). La SG a révolutionné l'amélioration des bovins car elle a augmenté la précision de sélection chez les jeunes animaux, rendant possible la sélection précoce et raccourcissant ainsi l'intervalle de génération. Enfin, elle a augmenté l'intensité de sélection en donnant une estimation de la valeur additive pour un plus grand nombre d'individus que la sélection classique basée sur les tests en descendance (Schefers et Weigel, 2012). L'application de la SG s'étend aujourd'hui aux autres espèces animales (porc, volailles, poissons, etc.) et aux plantes (maïs, blé, pommier, eucalyptus, palmier à huile, etc.) dont elle devrait aussi révolutionner l'amélioration (Desta et Ortiz, 2014).

III. B. 2. Précision de la sélection génomique

Comme pour la sélection classique, la précision de la sélection génomique est la corrélation entre la valeur additive vraie et son estimation, ici la GEBV. Dans les études utilisant des données réelles, elle est souvent estimée par validation croisée. Cette approche consiste à diviser un ensemble d'individus génotypés et ayant des observations (en général un phénotype) en plusieurs parties puis à utiliser successivement chaque partie comme une population de test dont les GEBV sont prédites par un modèle calibré avec les autres parties. On calcule alors dans chaque population de test la corrélation entre les GEBV et les valeurs disponibles des individus. Cependant, dans les études empiriques la valeur additive vraie des individus n'est pas connue. La corrélation calculée est alors la précision de prédiction (*predictive ability* ou *prediction accuracy*) et non pas de sélection. Elle mesure la capacité de la sélection génomique à prédire les valeurs observées, au lieu de la vraie valeur additive. Si les valeurs observées sont des phénotypes, on peut en déduire la précision de la sélection, qui vaut $r_{GEBV,A} = r_{GEBV,P} / h$, en supposant que les erreurs associées aux GEBV et au phénotype soient indépendantes, ce qui n'est probablement pas vérifié dans une validation croisée ou la population d'apprentissage et la population de test sont issues du même dispositif expérimental (Lorenz et al., 2011, p. 94).

La précision de la sélection génomique est affectée par l'apparentement entre populations d'apprentissage et de test, le nombre d'individus dans la population d'apprentissage, le déséquilibre de liaison entre les marqueurs et les gènes, la méthode statistique utilisée pour estimer les GEBV, l'héritabilité du caractère, la densité de marquage moléculaire, le type de marqueurs et l'architecture génétique du caractère (nombre de gènes et distribution de leurs effets) (Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014). Certains de ces facteurs sont traités plus en détails dans la suite de cette section.

III. B. 3. Modèles et méthodes statistiques de sélection génomique

Parmi les modèles et méthodes statistiques de SG on peut distinguer les approches qui estiment un effet additif associé à chaque marqueur et les approches qui donnent directement la GEBV (les modèles estimant des effets non additifs sont uniquement abordés dans la discussion de cette thèse).

Le nombre de marqueurs est normalement plus grand que le nombre d'individus génotypés et, par conséquent, l'estimation de l'effet des marqueurs par régression multiple avec la méthode des moindres carrés ordinaire n'est pas possible. Plusieurs alternatives existent : des méthodes utilisant le BLUP (RR-BLUP et GBLUP), des méthodes Bayésiennes (LASSO Bayésien, BayesA, BayesB, etc.) et des méthodes non paramètriques (*reproducing kernel Hilbert spaces* (RKHS), *random forest*, *support vector machine*, *neural nets*, etc.). Les méthodes non paramètriques ne sont pas traitées ici. Elles semblent surtout intéressantes pour les caractères faiblement additifs, pour lesquels elles peuvent donner des précisions plus élevées que les méthodes paramétriques (voir par exemple Gianola et van Kaam, 2008; Neves et al., 2012; González-Camacho et al., 2012; Ornella et al., 2014; Howard et al., 2014).

Pour les approches estimant un effet additif à chaque marqueur (notamment RR-BLUP, BRR, BayesA, BayesB, BayesCπ, BayesDπ et LASSO bayésien), le modèle de base est de la forme :

$$y = \mu + \boldsymbol{Z}m + e$$

avec y le vecteur des observations ($p \times 1$), μ la moyenne des observations, m le vecteur des effets aux marqueurs ($n \times 1$), $\boldsymbol{Z}$ une matrice d'incidence ($n \times p$) contenant le nombre de copies (0, 1 ou 2) de l'allèle le plus fréquent et e le vecteur des résidus ($p \times 1$), avec n marqueurs et p individus. On obtient alors une estimation de m, permettant d'obtenir la GEBV des individus candidats à la sélection en faisant la somme sur tous les marqueurs de leur effet additif respectif, c-à-d pour un individu i : $GEBV_i = \sum_{j=1}^{n} Z_{ij}\widehat{m}_j$.

Pour les approches donnant directement la GEBV (GBLUP), le modèle de base est de la forme :

$$y = \mu + g + e$$

avec g le vecteur ($p \times 1$) des GEBV, associé à une matrice d'apparentement moléculaire.

La SG estime, implicitement ou explicitement, l'effet de substitution α de chaque marqueur. Il faut cependant garder à l'esprit que, dans la mesure où l'on utilise essentiellement des marqueurs anonymes, ceux-ci ne possèdent pas réellement un effet sur le phénotype, mais l'analyse statistique leur attribut l'effet des gènes avec lesquels ils sont en déséquilibre de liaison.

Une grande diversité de méthodes statistiques ont été développées dans le cadre de la SG. Ceci s'explique notamment par le fait que les caractères quantitatifs d'intérêt reposent vraisemblablement sur différentes architectures génétiques, avec un nombre de gènes et une distribution de leurs effets variables. Ainsi, certaines méthodes statistiques de SG se basent sur une même variance additive (σ^2_m) pour tous les marqueurs (RR-BLUP, GBLUP, etc.), avec pour objectif de correspondre à des architectures génétiques proches du modèle infinitésimal (très grand nombre de gènes d'effets très faibles). D'autres méthodes (BayesA, BayesB, etc.) estiment des variances additives spécifiques à chaque marqueur. Ceci permet d'accorder une importance plus grande à certains marqueurs et devrait être adapté pour des caractères dont le déterminisme génétique implique aussi des gènes d'effets plus forts. L'expression « variance (additive) au marqueur » représente en fait un abus de langage, puisque σ^2_m diffère de la variance additive à un locus telle qu'elle est normalement définie en génétique quantitative (II. C). En réalité, deux interprétations peuvent être données à σ^2_m (Gianola et al., 2009). Dans la première, si on considère que l'effet d'un marqueur est estimé au cours d'un processus d'échantillonnage aléatoire dans une population d'effets possibles, σ^2_m correspond à la variance conceptuelle de la population d'effets possibles. La seconde interprétation s'applique aux méthodes bayésiennes. Dans celles-ci, σ^2_m représente l'incertitude autour de l'estimation de l'effet du marqueur. Par exemple, σ^2_m=0 indique un effet au marqueur estimé avec une parfaite certitude, sans impliqué que l'effet soit nul.

Les sections suivantes présentes les méthodes statistiques les plus couramment employées dans la SG. Plusieurs d'entre elles ont été appliquées au Chapitre V et leurs caractéristiques et propriétés ont été détaillées dans la partie « Matériel et méthodes ». On insistera particulièrement sur le GBLUP qui sert au Chapitre V et au Chapitre VI. Pour plus d'informations, on pourra se reporter aux articles qui traitent de la diversité des méthodes

statistiques de SG (Jannink et al., 2010; Lorenz et al., 2011; Heslot et al., 2012; de los Campos et al., 2013).

III. B. 3. a. RR-BLUP, BRR et BayesCπ

Le RR-BLUP ou *random regression* BLUP (Meuwissen et al., 2001) est une application du modèle mixte à la prédiction de l'effet (aléatoire) des marqueurs. Selon les auteurs il peut aussi être nommé *ridge regression* BLUP ou simplement BLUP. Tous les marqueurs ont la même variance, avec le vecteur des effets aux marqueurs $m \sim N(0, \sigma^2_m)$ et $e \sim N(0, \sigma^2_e)$. Les variances sont estimées par REML.

Le BRR ou *Bayesian random regression* (Pérez et al., 2010) est la version bayésienne du RR-BLUP. Il considère donc aussi que tous les marqueurs ont la même variance. Les méthodes bayésiennes utilisent les informations a priori fournies par l'utilisateur. Avec le BRR, ces informations portent sur σ^2_m et σ^2_e, qui ont une distribution a priori de type Chi² inverse, et sur m qui suit une loi normale $N(0, \sigma^2_m)$. Pérez et al. (2010) détaillent comment choisir les différents paramètres des distributions a priori (pour les distributions Chi² inverse, un degré de liberté et un paramètre d'échelle).

La méthode BayesCπ (Habier et al., 2011) est une extension de BRR considérant qu'une proportion π des marqueurs, estimée par le modèle, ont un effet nul. Pour les autres marqueurs, les règles de BRR s'appliquent. La distribution a priori de π est une loi bêta. Avec π=0, BayesCπ est identique à BRR.

La Figure 11 illustre la distribution a priori des effets aux marqueurs de BRR, de BayesCπ et d'autres méthodes bayésiennes (BLR, BayesA). Le Tableau 3 présente les propriétés et caractéristiques importantes du RR-BLUP, de BayesA et de BayesB.

III. B. 3. b. Bayesian LASSO regression

Le LASSO ou *least absolute shrinkage selection operator* est une méthode développée par Tibshirani (1996) pour estimer les paramètres de régressions linéaires. Il a été étendu sous une forme bayésienne par Park et Casella (2008), le *Bayesian LASSO regression* (BLR). Le BLR a été appliqué à la SG par de los Campos et al. (2009).

Avec le BLR, σ^2_e a une distribution a priori de type Chi² inverse, m a une distribution a priori normale avec une variance spécifique à chaque marqueur $\sigma^2_{m(i)} = \tau^2_i \times \sigma^2_e$, τ^2_i a une distribution a priori exponentielle de paramètre $\lambda^2/2$ et λ une distribution a priori de type gamma. Le BLR considère donc une variance spécifique à chaque marqueur. Pérez et al. (2010) détaillent comment choisir les différents paramètres des distributions a priori.

III. B. 3. c. BayesA, BayesB et BayesDπ

Les méthodes BayesA et BayesB font parties des méthodes proposées par Meuwissen et al. (2001). BayesA est proche de BLR dans le sens où elle considère une variance spécifique à chaque marqueur $\sigma^2_{m(i)}$. La distribution a priori de $\sigma^2_{m(i)}$ est une loi de Chi² inverse et celle de m une loi normale $N(0, \sigma^2_{m(i)})$. BayesB est basée sur BayesA, avec un

niveau de complexité supplémentaire, c-à-d qu'il considère qu'une proportion π des marqueurs ont un effet nul. Le paramètre π est spécifié par l'utilisateur et pour les autres marqueurs, les règles expliquées pour BayesA s'appliquent. Avec π=0, BayesB est donc identique à BayesA. L'estimation des $\sigma^2_{m(i)}$ et des m_i est faite par des méthodes de Monte-Carlo par chaînes de Markov, l'algorithme de Métropolis-Hastings et l'échantillonnage de Gibbs, respectivement.

La méthode BayesDπ a été proposée par Habier et al. (2011). Il s'agit d'une extension de BayesB dans laquelle le paramètre π est estimé par le modèle. Comme pour BayesCπ, la distribution a priori de π est une loi bêta.

III. B. 3. d. GBLUP

i) Principe

Le BLUP génomique ou GBLUP est basé sur le modèle mixte classique des évaluations génétiques (II. G), en remplaçant la matrice d'apparentement généalogique ***A*** par une matrice d'apparentement moléculaire ***G*** (VanRaden, 2007; Habier et al., 2007). La matrice ***A*** donne un apparentement attendu, en ignorant l'échantillonnage aléatoire des allèles parentaux à chaque locus au moment de la méiose (ségrégation mendélienne) (II. D). Au contraire, la matrice ***G*** estime l'apparentement réalisé en tenant compte de la ségrégation mendélienne. Grâce à cette information additionnelle elle peut donner des prédictions de valeurs génétiques plus précises que la matrice ***A***. Sous l'hypothèse de normalité des effets aux marqueurs, le GBLUP est équivalent au RR-BLUP.

*ii) Calcul de la matrice **G***

Plusieurs méthodes sont utilisées pour calculer la matrice d'apparentement réalisé ***G***. La méthode de Lynch améliorée par Li (Lynch, 1988; Li et al., 1993) utilise un indice de similarité appliqué à chaque locus et à partir duquel on calcule le coefficient de parenté. Pour un locus, si on considère deux individus *i* et *j* l'indice de similarité est :

$$S_{ij} = (I_{11} + I_{12} + I_{21} + I_{22}) / 4$$

où I_{xy} vaut 1 si l'allèle *x* de l'individu *i* est identique à l'allèle *y* de l'individu *j* et 0 sinon. S_{ij} peut donc prendre quatre valeurs : 0, 0.25, 0.5 et 1. Le coefficient de parenté moléculaire vaut :

$$f_{ij} = \frac{1}{q} \sum_{l=1}^{q} \left(S_{ij}\right)_l \qquad [20]$$

avec *q* le nombre de loci. D'une manière pratique, on obtient facilement les coefficients de parenté d'un ensemble d'individus par un calcul matriciel. En organisant les données moléculaires sous la forme d'une matrice (***Z***) contenant le nombre de copies de l'allèle présent à chaque locus de chaque individu (0, 1 ou 2), avec en colonne les données moléculaires par allèles et en ligne les individus, la matrice ***G*** des coefficients d'apparentement se calcule facilement :

$$\boldsymbol{G} = \boldsymbol{Z}\,{}^{t}(\boldsymbol{Z}) / 2q$$

avec ${}^{t}(\boldsymbol{Z})$ la transposée de ***Z***.

Cette méthode suppose que les allèles identiques (IBS) soient tous IBD, c'est-à-dire que chaque allèle était présent à l'origine en une seule copie dans la population des fondateurs (Eding et Meuwissen, 2001). Dans le cas contraire, f_{ij} est surestimé. Il est possible de corriger la formule [20] en tenant compte de la probabilité s_l qu'à un locus l un allèle de i soit IBS mais pas IBD à un allèle de j :

$$f_{ij} = \frac{1}{q} \sum_{l=1}^{q} \frac{(S_{ij})_l - s_l}{1 - s_l}.$$

Cependant, cette correction est difficile à mettre en œuvre car elle nécessite de connaitre le génotype de fondateurs qui ne sont pas apparentés entre eux.

Dans une autre méthode, le calcul des coefficients d'apparentement utilise des génotypes corrigés par les fréquences allèliques (VanRaden, 2007; Habier et al., 2007). Pour des marqueurs biallèliques, la matrice $\boldsymbol{G}$ correspondante vaut :

$$\boldsymbol{G} = \frac{\boldsymbol{X}\,\boldsymbol{t}(\boldsymbol{X})}{2\sum_{l=1}^{q} p_l(1-p_l)}$$

avec $\boldsymbol{X} = \boldsymbol{Z} - \boldsymbol{P}$, $\mathbf{Z}$ codée en 0, 1 ou 2 selon le génotype et $\boldsymbol{P}$ une matrice avec les individus en ligne, les loci en colonnes et chaque colonne l remplie de $2p_l$, c'est-à-dire du double de la fréquence de l'allèle le moins fréquent au locus l (ou de $2(p_l - 0.5)$ pour un codage de $\boldsymbol{Z}$ en –1, 0 ou 1). La division par $2\sum_{l=1}^{q} p_l(1-p_l)$ rend $\boldsymbol{G}$ analogue à $\boldsymbol{A}$. La soustraction de $\boldsymbol{P}$ à la matrice des génotypes donne plus de poids aux allèles rares qu'aux allèles fréquents.

Legarra (comm. pers.) propose une méthode pour étendre cet estimateur au cas des marqueurs multiallèliques :

$$\boldsymbol{G} = \frac{\boldsymbol{X}\,\boldsymbol{t}(\boldsymbol{X})}{2\sum_{l=1}^{q}(1-\sum_{k=1}^{n_l} p_{l_k}^2)}$$

avec n_l le nombre d'allèles du locus l et p_{l_k} la fréquence de l'allèle k du locus l.

Normalement, les fréquences allèliques sont celles de la population de fondateurs. Dans la pratique, celles-ci sont souvent inconnues et on utilise à la place les fréquences allèliques de la population étudiée. La matrice d'apparentement obtenue est parfois nommée $\boldsymbol{G_{OF}}$ (pour *observed allelic frequencies*).

Forni et al. (2011) proposent une méthode pour « normaliser » la matrice d'apparentement précédente de manière à ce que la valeur moyenne de sa diagonale soit de 1 :

$$\boldsymbol{G} = \frac{\boldsymbol{X}\,\boldsymbol{t}(\boldsymbol{X})}{trace(\boldsymbol{X}\,\boldsymbol{t}(\boldsymbol{X}))/n}$$

où n est le nombre d'individus. Cette méthode vise à obtenir des apparentements moléculaires présentant une compatibilité optimale avec les apparentements généalogiques, notamment dans le cas où les deux types d'apparentement doivent être combinés, par exemple pour analyser conjointement des individus génotypés et non génotypés. Dans le cas où la consanguinité n'est pas négligeable, on modifie la formule pour obtenir en moyenne sur la diagonale une valeur de $1 + F$, avec F la consanguinité généalogique moyenne de la population.

iii) Précision

Habier et al. (2007, 2010, 2013) ont étudié en détail les informations utilisées par le GBLUP pour prédire les GEBV des candidats à la sélection. Ils ont conclu que la précision du

GBLUP dépendait de l'apparentement additif entre les populations d'apprentissage et de sélection ainsi que du déséquilibre de liaison (DL) entre marqueurs et gènes au sein des individus considérés, qui peut se décomposer entre le DL dans la population de fondateurs et la coségrégation entre marqueurs et gènes depuis la population de fondateurs jusqu'aux individus considérés. La précision issue du DL est celle qui diminue le moins rapidement avec les générations de sélection sans recalibration du modèle.

La précision de sélection du GBLUP peut aussi se calculer avec la formule utilisant la PEV (éq. [14]).

III. B. 4. Effet du DL et de N_e sur la précision de la SG

La taille efficace (N_e) de la population et le déséquilibre de liaison (DL), deux notions étroitement liées (voir II. J), influencent fortement la précision de la SG, en interaction avec la densité de marquage (voir Figure 12 et section III. B. 5. b) et la taille de la population d'apprentissage (Heffner et al., 2009; Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014). La précision de la SG augmente quand N_e diminue. En effet, une N_e réduite implique un DL élevé, ce qui limite le nombre de marqueurs nécessaires pour avoir chaque gène lié à un marqueur. Par ailleurs, la densité de marquage et la taille de la population d'apprentissage doivent augmenter linéairement avec N_e pour maintenir la précision de la SG.

Quand on considère l'apparentement réalisé entre paires d'individus au sein d'une population, il existe une variabilité autour d'une valeur attendue. Le nombre efficace de loci d'une population (M_e) correspond au nombre de loci indépendants qui auraient donné la même variance d'apparentement réalisée que celle observée réellement (Goddard, 2009). Daetwyler et al. (2010) ont trouvé que la précision de la SG était fortement influencée par M_e, qui dépend de N_e et de la taille du génome en Morgan.

III. B. 5. Marquage moléculaire

III. B. 5. a. Type de marqueurs

Dans la publication de Meuwissen et al. (2001), la sélection génomique prédisait la valeur additive associée à des haplotypes composés de deux allèles de marqueurs multiallèliques adjacents. Solberg et al. (2008) ont comparé par simulation la sélection génomique avec des marqueurs SSR (multiallèliques) et SNP (biallèliques). Ils ont mis en évidence que pour un même nombre de marqueurs les SSR donnaient une précision plus élevée que les SNP et qu'il fallait 2 à 3 fois plus de SNP que de SSR pour parvenir à des précisions similaires.

Plusieurs études ont comparé l'utilisation de données de marqueurs pris individuellement ou rassemblés en haplotypes constitués d'allèles en phase de marqueurs adjacents (Calus et al., 2008, 2009; Solberg et al., 2008; Villumsen et al., 2009). Les conclusions obtenues étaient variables selon la densité de marquage, le type d'haplotypes (identiques par état ou par descendance, et avec un nombre plus ou moins grand de marqueurs), l'héritabilité du caractère et le nombre de générations de sélection sur marqueurs

sans recalibration du modèle. On peut considérer que l'utilisation de marqueurs individuels est une bonne option, en particulier lorsqu'il est possible d'avoir une forte densité de marquage. Ceci simplifie le travail en évitant l'étape de reconstitution des phases puis de conversion des données de marqueurs en données haplotypiques utilisables avec un modèle de SG.

III. B. 5. b. Densité de marquage

Plusieurs études ont porté sur l'effet de la densité de marquage moléculaire sur la précision de la sélection génomique (Calus et Veerkamp, 2007; Calus et al., 2008; Solberg et al., 2008; Meuwissen, 2009; de Roos et al., 2009). Elles ont montré que la précision augmentait avec le nombre de marqueurs avant d'atteindre un plateau. Par ailleurs, exprimer la densité de marquage en N_e par Morgan (N_e / M) facilite la comparaison de la précision entre espèces ou populations. Pour des candidats à la sélection apparentés à la population d'apprentissage (par exemple pour sélectionner parmi les descendants de la population d'apprentissage), une densité de marquage de 8 à 16 N_e / M permettrait d'obtenir une précision maximale. Compte tenu du lien entre N_e et DL, la densité de marquage peut aussi s'exprimer en termes de r^2 entre marqueurs adjacents. Meuwissen (2009) a trouvé que la précision augmentait linéairement avec le r^2 entre marqueurs adjacents. D'après Calus et Veerkamp (2007) un r^2 moyen de 0.15 serait suffisant pour un caractère avec une héritabilité de 0.5 et un r^2 de 0.2 pour un caractère avec une héritabilité faible (0.1).

III. B. 6. Définition de la population d'apprentissage

Dans la perspective de l'application pratique de la SG, la population d'apprentissage doit correspondre à une phase opérationnelle du programme (plutôt qu'à des individus installés dans des essais spécifiques), afin de pouvoir calibrer un modèle de SG utilisable directement chez les candidats à la sélection, qui seront généralement des descendants ou des collatéraux de la population d'apprentissage.

La précision de la SG augmente avec la taille de la population d'apprentissage, qui apparait plus souvent comme un facteur limitant de la SG que le nombre de marqueurs (Jannink et al., 2010; Lorenz et al., 2011; Grattapaglia, 2014). Par ailleurs, l'apparentement entre la population d'apprentissage et les candidats à la sélection est un facteur capital de la précision de la SG, qui est d'autant plus forte que les deux populations sont apparentées (Pszczola et al., 2012; Ly et al., 2013; Daetwyler et al., 2013; Gowda et al., 2014). Muir (2007) a montré qu'il était avantageux d'avoir une population d'apprentissage composée d'individus de plusieurs générations. Il a comparé l'évolution de la précision de la SG après avoir calibré le modèle avec 2 048 individus répartis sur deux générations ou sur quatre générations. Il a constaté qu'avec quatre générations représentées dans la population d'apprentissage la précision se maintenait plus longtemps dans les générations suivant la calibration du modèle (Figure 13).

Rincent et al. (2012) ont développé une méthode, nommée CDmean, visant à concevoir une population d'apprentissage qui maximise la précision chez les candidats à la

sélection. Elle est applicable lorsque l'on dispose d'un ensemble d'individus génotypés et que l'on souhaite en tirer un sous ensemble qui servira de population d'apprentissage pour prédire la GEBV des individus restants. Le critère d'optimisation est le coefficient de détermination généralisé, qui correspond au carré de la corrélation attendue entre la vraie valeur et la valeur estimée du contraste des valeurs génétiques, c-à-d la fiabilité (voir p. 17) attendue des contrastes entre les individus non phénotypés et la moyenne de la population.

III. B. 7. La sélection génomique pour la performance d'hybrides

Plusieurs études ont porté sur l'utilisation de la SG dans le contexte de l'amélioration génétique pour la production d'hybrides, chez les plantes et les animaux (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Kinghorn et al., 2010; Technow et al., 2012; Reif et al., 2013; Zeng et al., 2013; Xu et al., 2014). Elles ont notamment traité le type de modèle (purement additif ou avec des effets additifs et de dominance, avec des effets aux marqueurs spécifiques ou communs aux populations parentales) et le jeu de données à utiliser pour la calibration (individus hybrides en distinguant ou non l'origine parentale des allèles, ou parents).

Kinghorn et al. (2010) ont simulé pendant 40 générations un schéma d'amélioration hybride entre deux lignées parentales, en considérant un caractère avec de l'hétérosis due à des effets de dominance au niveau des gènes. Ils ont appliqué la SG pour sélectionner des parents dans le but d'augmenter la valeur des hybrides. Ils ont comparé trois stratégies de SG nommées *reciprocal recurrent genomic selection* (RRGS), *crossbred genomic selection* (XBGS) et *within line genomic selection* (WLGS), avec des modèles purement additifs. La calibration dans la RRGS utilise le phénotype des hybrides et leur génotype, en distinguant à chaque marqueur la lignée parentale d'origine des allèles. Le modèle estime alors deux effets par marqueur, chacun spécifique à une lignée parentale. Avec la XBGS, la calibration utilise aussi le phénotype et le génotype des hybrides, mais considère que les marqueurs n'ont qu'un seul effet, identique pour les deux lignées parentales. Dans la WLGS, le modèle estime des effets aux marqueurs spécifiques aux populations parentales (comme la RRGS) mais à partir du phénotype et du génotype des parents. La RRGS a abouti a un gain génétique très fort, la WLGS un gain faible et la XBGS un gain intermédiaire. La mauvaise performance de la WLGS traduit le fait qu'en présence de dominance l'effet moyen intrapopulation d'un allèle (tel qu'estimé par ce modèle) diffère de son effet moyen interpopulation (cf III. A. 2. a). La WLGS n'est donc pas pertinente pour prédire l'aptitude à la combinaison hybride des parents, contrairement à la RRGS et à la XBGS qui estiment des effets aux marqueurs interpopulation. Enfin, la supériorité de la RRGS par rapport à la XBGS traduit le fait que la différentiation génétique entre les deux lignées parentales était suffisamment forte pour rendre préférable l'estimation d'effets aux marqueurs spécifiques à chaque lignée.

Ibánez-Escriche et al. (2009) ont simulé un schéma d'amélioration hybride entre deux populations plus ou moins distantes, avec un caractère ayant un déterminisme additif. Ils ont comparé des modèles de SG avec des effets aux marqueurs spécifiques ou communs aux populations parentales. Ils ont conclu que l'estimation d'effets spécifiques aux populations parentales donnait de meilleures précisions de SG lorsque le nombre de marqueurs était faible

(500), le nombre d'individus important et les populations parentales peu ou pas du tout apparentées. Technow et al. (2012) ont simulé un schéma d'amélioration hybride chez le maïs, avec un caractère sous le contrôle d'effets additifs et de dominance. Ils ont montré que l'utilisation d'un modèle de SG avec des effets aux marqueurs spécifiques aux populations parentales était surtout intéressante lorsque le déséquilibre de liaison interpopulation était faible. Ils ont aussi trouvé que l'utilisation d'un terme de dominance dans le modèle était utile lorsque la part de la variance de dominance dans la variance génétique totale entre hybrides était forte. Zeng et al. (2013) ont obtenu des résultats similaires avec des simulations d'un programme hybride entre deux populations. Ils montré que plus la part de dominance dans le déterminisme génétique du caractère était importante, plus la réponse à la sélection du modèle avec dominance dépassait celle du modèle additif. En l'absence de dominance, les deux modèles avaient des performances similaires.

L'utilisation de la SG pour augmenter de la vigueur hybride résultant de l'interaction multiplicative entre deux composantes antagonistes et additives n'a pas été étudiée. Elle le sera dans le Chapitre VI avec le palmier à huile, grâce à une simulation portant sur la production de régimes et ses deux composantes, le poids moyen et le nombre de régimes.

III. B. 8. La sélection génomique pour les espèces pérennes et le palmier à huile

L'utilisation de la SG pour les espèces pérennes a été traitée en détail dans deux articles de revue (Grattapaglia, 2014; Isik, 2014). Pour le palmier à huile, une seule étude a été publiée (Wong et Bernardo, 2008). Ces publications sont reprises dans les introductions des articles constituant le Chapitre V et le Chapitre VI.

La principale difficulté commune à l'amélioration des espèces pérennes est la longueur de l'intervalle de génération. Par ailleurs, les évaluations au champ sont généralement coûteuses et complexes à réaliser, nécessitant de grandes superficies et une main d'œuvre importante pendant de nombreuses années (5-20 ans). Le principal intérêt de la SG serait donc de raccourcir l'intervalle de génération en permettant la sélection d'individus sans avoir à les évaluer. Selon les espèces, la SG pourrait aussi augmenter l'intensité de sélection et augmenter la précision de sélection chez les individus évalués. Chez le palmier à huile, les dispositifs expérimentaux permettent d'avoir des estimations fiables des AGC parentales avec le modèle mixte traditionnel, sans marqueurs (précision de sélection estimée à 0.9 au Chapitre IV). Les paramètres sur lesquels on essayera de faire agir la SG pour accroitre le rythme du progrès génétique sont donc l'intervalle de génération (actuellement 20 ans) et la faible intensité de sélection (actuellement moins de 200 individus testés en croisement par groupe parental et par génération, parmi les quelques milliers disponibles). Par rapport à beaucoup d'espèces pérennes, le palmier à huile se distingue par un petit nombre d'individus évalués (petites populations d'apprentissage pour la SG) et par un déséquilibre de liaison fort.

Wong et Bernardo (2008) ont comparé par simulations la sélection phénotypique classique, la MARS (sélection récurrente assistée par marqueurs, basée sur la détection de QTL) et la SG en termes de réponse à la sélection chez le palmier à huile. Ils ont simulé une petite population (60 à 140 individus) issue de l'autofécondation d'un croisement entre lignées

pures. Dans la première génération, ces individus ont été croisés avec un testeur puis sélectionnés sur leur valeur en croisement. Les meilleurs sont ensuite recombinés entre eux pour produire la génération suivante. Pour la sélection phénotypique cette procédure est répétée dans le cycle suivant. Pour la MARS et la SG, cette population initiale est génotypée et utilisée pour estimer des effets associés aux marqueurs. Pendant les trois cycles suivants (soit un total de quatre cycles, équivalents en nombre d'années aux deux cycles d'amélioration classique), la sélection est faite uniquement sur les marqueurs. Dans la MARS, les auteurs détectent parmi 100 marqueurs ceux qui sont significativement associés à des QTL. Pour la SG, 140 marqueurs sont utilisés. Finalement, la SG donne les meilleurs résultats en termes de gain génétique par unité de coût et par an (Figure 14). Cependant, ces simulations sont basées sur des hypothèses fortes, qui font que la population utilisée pour calibrer le modèle de SG est très différente des populations parentales qui existent dans les programmes d'amélioration du palmier à huile, beaucoup plus complexes (cf III. A. 3. c).

CONCEPTION DES ETUDES REALISEES DANS LA THESE

Les études réalisées au cours de cette thèse ont été planifiées sur la base des caractéristiques biologiques du palmier à huile, de son schéma d'amélioration, des données disponibles et des informations de la littérature présentées dans la revue bibliographique.

Les données disponibles étaient les observations phénotypiques faites dans un dispositif expérimental visant à tester en croisement 146 individus A et 156 individus B appartenant aux populations d'amélioration de PalmElit et de ses partenaires (SOCFINDO, CRAPP), acquises avant la thèse. Ces individus ont été génotypés avec 265 SSR au CIRAD, une tâche qui s'est achevée au début de la thèse. Mon travail sur les données moléculaires s'est limité à vérifier leur qualité, repérer les données suspectes (notamment sur la base d'incohérences par rapport à la généalogie) puis à effectuer des relectures. Ces données phénotypiques et moléculaires ont été utilisées pour caractériser les populations d'amélioration (Chapitre IV) et pour l'étude empirique de la précision de la SG (Chapitre V). Afin d'étendre l'étude au-delà de ce qui pouvait être traité avec les données expérimentales et de répondre à certaines questions soulevées par les résultats empiriques, des simulations ont été réalisées (Chapitre VI).

La caractérisation des populations d'amélioration (Chapitre IV) avait plusieurs objectifs. Elle visait tout d'abord à estimer les paramètres génétiques susceptibles d'influencer la précision de la SG (Chapitre V). Les paramètres retenus ont été ceux présentés dans la revue bibliographique : la taille efficace (N_e) des populations, l'héritabilité des caractères (h^2) et l'apparentement entre les individus au sein de chaque groupe parental. Elle a aussi servi à estimer les AGC des parents, utilisées pour calibrer le modèle de SG et pour en mesurer la précision (Chapitre V). Enfin, certains paramètres génétiques (différentiation génétique (F_{ST}), variances génétiques additives pour le nombre de régimes (NR) et leur poids moyen (PM) et corrélation génétique additive entre NR et PM) ont servi à calibrer de manière pertinente les simulations du Chapitre VI, en plus d'autres paramètres obtenus dans la littérature.

La revue bibliographique a permis d'identifier les facteurs à faire varier dans l'étude empirique de la précision de la SG (Chapitre V), car susceptibles d'influencer les résultats : méthode statistique d'analyse du modèle de SG, caractéristique des populations d'apprentissage et de validation (apparentement entre elles et optimisation de la population d'apprentissage), population et caractères d'intérêt (compte tenu de la variabilité de h^2 et en supposant une diversité d'architecture génétique entre caractères et populations). Enfin, la revue bibliographique a aidé à définir l'étude par simulation, en particulier avec la publication de Kinghorn et al. (2010) qui présente la sélection génomique récurrente réciproque.

CHAPITRE IV. ETUDE PRELIMINAIRE : CARACTERISATION DES POPULATIONS D'AMELIORATION

Ce chapitre est une étude préliminaire visant à préciser les caractéristiques génétiques des populations de travail : apparentement généalogique et moléculaire, taille efficace, différentiation génétique entre populations, paramètres génétiques (variances des caractères considérés, corrélations génétiques additives et résiduelles entre nombre et poids de régime, etc.) et AGC.

IV. A. Matériel et méthodes

Les données phénotypiques disponibles sont celles de l'ensemble de tests en descendance installés à Aek Loba, en Indonésie (Sumatra). Il s'agit de 26 essais plantés entre 1995 et 2000 (Figure 15) visant à évaluer des croisements hybrides A × B pour toutes les composantes du rendement. Le but était d'estimer l'AGC des candidats à la sélection du second cycle d'amélioration du palmier à huile. Un des essais (BBGT28) est constitué de croisements entre les parents de ces candidats, c-à-d entre les individus sélectionnés à l'issue du premier cycle d'amélioration.

Les données moléculaires ont été acquises sur la majorité des parents des croisements des 26 essais (à l'exception de quelques individus dont l'ADN n'était pas disponible) et sur quelques uns de leurs ancêtres.

Le groupe A était essentiellement constitué de Deli et de quelques Angola. Le groupe B était composé de plusieurs populations africaines. Les tests en descendance ont fait l'objet d'une analyse globale en prenant en compte les parents de toutes les populations (y compris les Angola exclus des autres études) afin de garder un dispositif le plus équilibré possible et de garantir les connexions entre essais et entre parents, ainsi que la meilleure estimation possible des effets non génétiques (essais, blocs, etc.). Les pédigrées sont donnés au Chapitre V dans la Figure S1 pour les Deli et dans la Figure S2 pour le groupe B. Les individus en couleur (131 Deli et 131 du groupe B) ont été à la fois génotypés et testés en croisement, et ils représentent le matériel d'étude de la validation empirique de la sélection génomique (Chapitre V). Quelques individus présents dans le pédigrée mais non testés en croisement ont aussi été génotypés et ont été inclus dans l'étude sur la reconstruction du pédigrée des Deli et dans le calcul des tailles efficaces avec le logiciel LDNE.

IV. A. 1. Populations d'améliorations et données moléculaires

Les tests sur descendance considérés dans cette thèse impliquent 146 individus A et 156 individus B. Le groupe A compte 135 Deli parmi lesquels 131 ont été génotypés. Dans le groupe B, 131 individus ont été génotypés, répartis entre 94 La Mé, 24 Yangambi, 4 La Mé × Yangambi, 7 La Mé × Sibiti et 2 WAIFOR. Les génotypages ont été faits avec 265 marqueurs microsatellites (SSR), décrits dans Billotte et al. (2005) et Tranbarger et al. (2012). 220 SSR se sont révélés polymorphes chez les Deli et 260 dans le groupe B, donnant une densité de marqueurs polymorphes d'un SSR par 7.9 cM chez les Deli et un SSR par 6.7 cM chez le groupe B, en considérant une longueur de génome de 1 743 cM (Billotte et al., 2005). Les SSR polymorphes possédaient en moyenne 2.7 ± 0.8 (ET) allèles chez les Deli et 6.2 ± 2.2 dans le groupe B. Les relations connues de parenté ont été utilisées pour vérifier la qualité des génotypages et aider à corriger les erreurs. Elles ont aussi été utilisées pour imputer des données moléculaires manquantes (par exemple, pour déduire le génotype à un locus d'un individu issu de parents homozygotes). Après cette opération, le pourcentage de données moléculaires manquantes était très faible, à 1.74% chez les Deli et 2.90% dans le groupe B.

IV. A. 2. Apparentement généalogique et moléculaire

L'estimation de l'apparentement généalogique nécessite un enregistrement complet du pédigrée.

Le pédigrée des populations africaines est assez bien connu. Pour les La Mé, et même les Yangambi dont l'histoire est plus longue, la généalogie remonte en général jusqu'aux fondateurs (cf pédigrée dans la Figure S2 du Chapitre V). L'histoire de la population Deli est globalement connue mais le pédigrée précis n'est disponible que sur les dernières générations. Il s'agit donc d'un cas d'étude intéressant pour les méthodes de reconstruction de pédigrée utilisant des données moléculaires. Par exemple, la méthode de Fernández et Toro (2006) fournit le pédigrée le plus vraisemblable pour un ensemble d'individus contemporains compte tenu de leur génotype, grâce à un algorithme d'optimisation, le recuit simulé (*simulated annealing*). Celui-ci est utilisé pour identifier un pédigrée qui maximise la corrélation entre matrices d'apparentement moléculaire et généalogique. La méthode de Fernández et Toro (2006) a été implémentée dans le logiciel MOLCOANC mais elle est limitée aux espèces dioïques et à un nombre constant d'individus par génération dans le pédigrée à reconstruire. Dans le cadre de cette thèse, la méthode de Fernández et Toro (2006) a été étendue aux modes de reproduction monoïque et hermaphrodite avec possibilité d'autofécondation, et a été rendue plus flexible, en donnant la possibilité de spécifier plusieurs paramètres : (i) un nombre d'individus spécifique à chaque génération du pédigrée à reconstruire, (ii) la première génération à partir de laquelle les autofécondations sont possibles et (iii) une matrice d'apparentement prédéfinie entre les fondateurs. La méthode étendue a été validée avec des données simulées et avec les données réelles de la population Yangambi. Elle a ensuite été appliquée à la population Deli. Le détail de la méthode est décrit dans la publication correspondante (Cros, Sánchez, et al., 2014) présentée à l'Annexe 3.

La parenté moléculaire a été calculée selon Lynch (1988) et Li et al. (1993) (voir III. B. 3. d.ii)). Une représentation graphique sous forme de carte de chaleur et de dendrogramme a été obtenue avec la fonction *heatmap.*2 du package *gplots* de R.

IV. A. 3. Taille efficace

La taille efficace (N_e) de la population Deli, du groupe B et de la population La Mé a été calculée avec deux méthodes : le DL, grâce au logiciel LDNE (Waples et Do, 2008), et le pédigrée grâce au logiciel ENDOG (Gutiérrez et Goyache, 2005). Les calculs ont été faits à partir des données généalogiques ou moléculaires de 143 Deli, 136 individus du groupe B et 98 La Mé, incluant les individus utilisés pour la validation empirique de la sélection génomique (Chapitre V).

LDNE estime N_{eC} par la méthode de Hill (1981) corrigée par Waples (2006) (II. J). Pour chaque paire de loci *i* et *j* ayant k_i et k_j allèles, respectivement, il calcule $r^2_{\hat{\Delta}_{AB}}$ pour chacune des $k_i \times k_j$ combinaisons possibles d'allèles A et B présents aux loci *i* et *j*, respectivement. Un $r^2_{\hat{\Delta}_{AB}}$ moyen pondéré est calculé sur l'ensemble des paires de loci en tenant compte, pour chaque paire de loci, du nombre d'allèles et du nombre d'individus ayant des données génotypiques. Pour chaque population on a défini trois lots de 16 SSR polymorphes, en choisissant chaque SSR sur un groupe de liaison différent, aléatoirement. N_{eC} a ensuite été calculée avec LDNE pour chaque lot de 16 SSR. Le logiciel a été paramétré de manière à tenir compte de tous les allèles, quelque soit leur fréquence dans l'échantillon, et en faisant l'hypothèse d'une reproduction aléatoire dans la population.

ENDOG applique la méthode de Gutiérrez et al. (2008, 2009) et Cervantes et al. (2011) pour calculer N_{eC} et N_{eP} réalisées. Pour les individus du groupe B et La Mé il a été utilisé avec leur pédigrée connu (Figure S2 du Chapitre V) et pour la population Deli à partir du pédigrée reconstruit avec MOLCOANC.

La publication présentée à l'Annexe 3 (Cros, Sánchez, et al., 2014) détaille le calcul de N_{eC} pour les Deli avec LDNE et de N_{eC} et N_{eP} pour les Deli et Yangambi avec ENDOG.

IV. A. 4. Différentiation génétique entre populations

Les F-Statistiques de Wright (1931) correspondent à trois indices de fixation, le F_{IT}, le F_{IS} et le F_{ST}, calculés à partir des proportions d'hétérozygotes observées et attendues d'après le modèle de Hardy-Weinberg. Le F_{ST} mesure la différenciation génétique entre populations, il vaut :

$$F_{ST} = 1 - \frac{H_S}{H_T}$$

avec H_S la probabilité de tirer deux allèles différents dans deux individus de la même sous-population et H_T dans deux individus de sous-populations différentes. Le F_{ST} indique un déficit en hétérozygotes dû à la non-panmixie entre sous-populations. Un $F_{ST} = 0$ indique l'absence de variation entre sous-populations (par ex. migration libre), un $F_{ST} > 0$ indique une différenciation entre sous populations et un $F_{ST} = 1$ indique que chaque sous-population est

fixée pour l'un ou l'autre des allèles présents (absence de migration). Le F_{ST} se définit aussi en termes de consanguinité des populations relative à la consanguinité totale. Il mesure donc l'effet de la subdivision sur la consanguinité :

$$F_{ST} = \frac{Q_S - Q_T}{1 - Q_T}$$

avec Q_S (= 1 – H_S) la probabilité de tirer deux allèles identiques dans deux individus de la même sous-population et Q_T dans deux individus de sous-populations différentes.

Weir et Cockerham (1984) ont développé des estimateurs non biaisés des indices de fixation de Wright. L'estimateur θ_{WC} du F_{ST} vaut :

$$\theta_{WC} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_c^2}$$

avec σ^2_a la variance des fréquences allèliques entre les populations, σ^2_b la variance entre individus à l'intérieur des sous populations et σ^2_c la variance entre gamètes à l'intérieur des individus. La variance de θ_{WC} n'est minimale que lorsque le F_{ST} est supérieur à 0.10 (http://kimura.univ-montp2.fr/genetix/FST.htm).

Le F_{ST} a été mesuré dans la perspective de bien calibrer les simulations (0). Comme celles-ci portaient uniquement sur les populations Deli et La Mé, le F_{ST} a été mesuré seulement entre elles. Les données moléculaires utilisées sont celles de 220 SSR passés sur 143 Deli et 98 La Mé, incluant les individus utilisés pour la validation empirique de la sélection génomique (Chapitre V). Le calcul de θ_{WC} a été fait avec la fonction *fastDivPart* du package *diveRsity* (Keenan et al., 2013) de R.

IV. A. 5. Paramètres génétiques et BLUP

Les paramètres génétiques d'intérêt dans les populations parentales sont les variances (d'AGC, d'ASC, résiduelle et phénotypique) et l'héritabilité au sens strict des caractères considérés, les corrélations génétiques additives et résiduelles entre nombre et poids de régime et la précision des BLUP (AGC et ASC). Ces valeurs ont été obtenues par un modèle mixte traditionnel, c-à-d utilisant le pédigrée des parents A et B testés en croisement et les données phénotypiques des hybrides A × B observés en essais.

Le plan de croisements suivi pour l'installation des tests sur descendance était un plan factoriel incomplet avec 478 croisements entre 146 individus A et 156 B. Dans le groupe A, 135 individus étaient des Deli, impliqués dans 445 croisements, et les 11 autres étaient des Angola. Les croisements effectués représentaient seulement 2.1% des croisements possibles. Les croisements étaient évalués dans 26 essais plantés entre 1995 et 2000. Les essais ont été installés selon des dispositifs en RCBD avec cinq ou six répétitions ou en lattices de rang quatre ou cinq (c-à-d avec un nombre de répétitions [blocs complets] égal au rang + 1, un nombre de blocs incomplets égal au rang et un nombre de croisements égal au rang²).

On utilise ici les données de dix caractères liés à la production d'huile de palme : le poids total de régimes produits annuellement (PR), le nombre annuel de régimes (NR), le poids moyen annuel des régimes (PM), le pourcentage d'huile dans les régimes (%HR), de fruits dans les régimes (%FR), de pulpe dans les fruits (%PF), d'amande dans les fruits (%AF) et d'huile dans la pulpe (%HP), le poids moyen des fruits (PF) et le nombre de fruits par

régime (NF). La production de régimes a été mesurée à l'âge adulte, c-à-d de 6 à 11 ans, sur 30 872 palmiers de type tenera (type commercial) avec une observation tous les dix jours. PR est la production de régimes exprimée en kilogrammes par palmier et par an, NR est le nombre total de régimes récolté par palmier et par an et PM le poids moyen annuel des régimes récoltés. Les autres caractères, qui sont liés à la qualité des régimes (ou taux d'extraction, %HR) ont été mesurés sur 21 525 palmiers, à raison d'un échantillon d'au moins 24 palmiers par croisement, observés entre cinq et six ans.

L'AGC des parents et l'ASC des croisements ont été obtenues grâce à un modèle mixte classique, basé sur le modèle de Stuber et Cockerham (1966) (voir équation [18]), similaire à celui appliqué par Purba et al. (2001). Il est détaillé dans Cros et al. (2014), au Chapitre V (voir Appendix de la publication, p. 71). La précision des AGC et des ASC est calculée avec les formules données dans la section III. A. 2. b.

Pour NR et PM, une analyse conjointe a été réalisée avec un modèle bivarié, bien que cela n'ait pas été précisé dans Cros et al. (2014). Ceci permet d'estimer les corrélations additives et résiduelles entre ces deux caractères et d'améliorer la précision des BLUP. Les corrélations sont obtenues à partir de la covariance et des variances données par l'analyse du modèle mixte, en appliquant la formule : $r_{BN,PM} = \sigma_{BN,PM} / (\sigma_{BN} \sigma_{PM})$.

h^2 est calculée selon la formule donnée au III. A. 2. d, en faisant le rapport entre σ_{agc}^{2AB} et la variance phénotypique entre croisements.

IV. B. Résultats

IV. B. 1. Apparentement généalogique et moléculaire

L'adaptation du logiciel MOLCOANC a permis d'obtenir un pédigrée reliant tous les individus Deli aux quatre fondateurs de 1848. Ce pédigrée reconstruit a ensuite été utilisé pour estimer N_e réalisée de la population avec ENDOG.

La parenté moléculaire était en moyenne de 0.58 chez les Deli, avec une gamme allant de 0.42 à 0.96, et de 0.39 (0.12 – 0.92) dans le groupe B. Les représentations graphiques des matrices de parenté moléculaire sont données en Figure 16 pour les Deli et en Figure 17 pour le groupe B. On note que les valeurs hautes de parenté sont similaires entre le groupe B et la population Deli mais que les valeurs basses sont beaucoup plus faibles dans le groupe B, indiquant que celui-ci est plus fortement structuré. Ceci traduit la présence au sein du groupe B de plusieurs populations qui ont été assez peu mélangées, en particulier les La Mé et les Yangambi.

Au moment où la thèse a démarré les individus ont été assignés aux différentes populations selon les connaissances des sélectionneurs. On constate que ceci correspond bien aux résultats fournis par les données moléculaires. Par exemple, le dendrogramme associé à la matrice de parenté moléculaire indique trois grands ensembles dans le groupe B, l'un constitué des 24 individus Yangambi et des 2 du Nigéria, un autre de 85 La Mé et un troisième, intermédiaire, composé de 9 La Mé et des 11 La Mé × Sibiti et La Mé × Yangambi.

Dans le détail on peut noter que, si les sélectionneurs considèrent que les quatre palmiers illégitimes PO4100P, PO4098P, PO4964T et PO4096P appartiennent à la population Yangambi, ils sortent ici plus proches des deux palmiers du Nigéria (BB213P et BB227P) que des Yangambi. Ils pourraient donc être d'origine plus complexe, tel qu'un croisement entre Nigéria et Yangambi. L'arbre BB5197T, indiqué d'une origine « Lisombé-Kinshasa ? » dans les registres des sélectionneurs s'intègre ici parfaitement aux Yangambi, dans lesquels il a donc été rangé pour la thèse. On remarque aussi que le La Mé LM9T apparait dans le groupe intermédiaire et qu'il pourrait éventuellement être issu d'un croisement entre un La Mé et un individu d'Afrique Centrale. La majorité des individus La Mé illégitimes (c-à-d PO5491T, PO5488T et les individus issus du croisement noté LM9T × M_FR12 dans le pédigrée) apparaissent dans le groupe intermédiaire et pourraient donc être issus de croisements La Mé × Afrique Centrale. Seul PO3636P apparait clairement dans les La Mé, indiquant qu'il n'est certainement pas un plein-frère de PO5491T. Ces quelques incertitudes ont éventuellement pu affecter légèrement les résultats des calculs de N_e chez les La Mé et du F_{ST} entre Deli et La Mé, avec l'attribution possible de quelques individus dans la population La Mé alors qu'ils seraient en réalité des croisements entre des La Mé et d'autres populations africaines. Ceci ne peut par contre pas affecter les résultats de la sélection génomique puisque le groupe B est traité dans son ensemble, sans faire de distinction entre les populations qui le composent.

IV. B. 2. Taille efficace

Les N_{eC} obtenues à partir du déséquilibre de liaison (DL) sont assez proches entre les populations (Deli, La Mé et groupe B). Elles varient entre 3 et 5, et sont donc très faibles (Figure 18).

Les *N*e réalisées obtenues à partir du pédigrée sont données dans la Figure 19. Pour toutes les populations les N_{eC} réalisées sont proches des N_{eC} obtenues avec le DL.

IV. B. 3. Différentiation génétique entre populations

Le F_{ST} obtenu entre les populations Deli et La Mé est de 0.474, avec un intervalle de confiance à 95% de 0.467 à 0.488.

IV. B. 4. Paramètres génétiques et BLUP

Les estimations des variances additives interpopulations dans les groupes parentaux A et B et des variances de dominance dans la population hybride sont fournies dans le Tableau 4. Les héritabilités (h^2) et la fiabilité des AGC sont présentées dans la Figure 20. h^2 est plus élevée dans le groupe B que dans le groupe A. Ceci est certainement lié à l'histoire des deux groupes : le groupe A, en étant essentiellement composé de la population Deli, se distingue du groupe B par une base génétique plus étroite et par un plus grand nombre de générations de sélection, de croisements entre apparentés et de dérive génétique, avec comme conséquence

attendue une réduction de la variance additive. L'écart entre les deux groupes est plus ou moins fort selon le caractère. Par exemple dans le groupe A, h^2 est plus grande pour NR que PM mais dans le groupe B l'inverse s'observe. A nouveau, ceci pourrait être lié à l'histoire des deux groupes chez qui, suite au hasard (dérive génétique) ou à l'effet de la sélection menée de manière indépendante au sein de chaque groupe, la réduction de la variance additive dans le groupe A par rapport au groupe B aurait pu se retrouver plus ou moins marquée selon le caractère.

On observe que la variance des AGC au sein des familles de plein-frères est systématiquement plus faible chez les Deli que dans le groupe B (entre 16% et 74% plus faible selon le caractère, voir Tableau 5).

La fiabilité des AGC (c-à-d leur précision élevée au carré) est en moyenne de 0.7 et elle est beaucoup moins variable que h^2. Elle est généralement plus forte dans le groupe A que dans le groupe B. L'ampleur de la supériorité de la fiabilité des AGC par rapport à h^2 traduit la quantité d'information apportée par la prise en compte des apparentements dans l'analyse (Walsh, 2013, p 8). Cette augmentation est beaucoup plus grande dans le groupe A (en moyenne +0.42) que dans le groupe B (+0.26). Comme le nombre d'individus dans les deux groupes parentaux est sensiblement le même, ceci est certainement lié à la plus grande structuration qui existe dans le groupe B : le fait que le groupe B soit constitué de populations qui ont été peu brassées a pour conséquence que les individus de ce groupe ont moins de collatéraux pour contribuer à l'estimation de leur AGC que les individus du groupe A. Dans ces conditions, la prise en compte des apparentements est plus profitable au groupe A qu'au groupe B.

Dans la suite de la thèse, la sélection génomique sera évaluée notamment par sa précision. Pour faire une comparaison pertinente avec la sélection classique (c-à-d le modèle mixte tel qu'il a été appliqué dans cette partie), nous avons calculé la précision moyenne de l'AGC des 131 Deli et des 131 individus du groupe B qui serviront à l'étude empirique de la sélection génomique, pour les huit caractères concernés (Tableau 6). La précision des AGC est globalement élevée, autour de 0.90. Bien que l'écart soit faible, elle est significativement plus forte dans le groupe B (en moyenne 0.91) que dans le groupe A (0.87).

A partir des variances et des covariances, on calcule que la corrélation génétique additive entre NR et PM est très forte puisqu'elle atteint –0.99 (erreur type ± 0.04) dans le groupe A et –0.94 (± 0.02) dans le groupe B. En comparaison, la corrélation résiduelle est faible (–0.16, ± 0.003).

La fiabilité des ASC est beaucoup moins forte que celle des AGC, puisqu'elle est en moyenne de 0.3 (Figure 21). Par ailleurs, elle est calculée sur les croisements qui ont effectivement été observés, et qui représentent un pourcentage très faible des croisements possibles. Si l'objectif des sélectionneurs était de valoriser les ASC il faudrait se pencher sur les ASC de tous les croisements et, pour le cas général des croisements qui n'ont pas été réalisés, la précision des ASC serait encore plus faible.

La Figure 22 donne le ratio entre la variance des ASC et la variance génétique entre croisements $\sigma^{2^{AB}}_{agc_A} + \sigma^{2^{AB}}_{agc_B} + \sigma^{2^{AB}}_{asc}$ (voir III. A. 2. a), et montre que la part des effets de dominance dans la variance génétique totale entre croisements est faible (ratio <15%).

IV. C. Discussion

Le pédigrée reconstruit des Deli est apparu intéressant pour estimer les N_e réalisées. L'utilisation du pédigrée reconstruit des Deli à la place du pédigrée connu dans le modèle mixte classique (pas montré) a eu un effet insignifiant sur l'AGC des individus testés en croisement et sur la vraisemblance du modèle. On suppose que, bien que le pédigrée connu des Deli ne renvoie pas jusqu'aux quatre fondateurs de 1848, le nombre de générations et / ou le nombre de descendants observés en essai par parent serait suffisamment élevée pour obtenir une estimation satisfaisante des AGC. Finalement, pour tous les calculs faisant appel au pédigrée à l'exception de l'estimation des N_e réalisées (c-à-d pour l'estimation des paramètres génétiques et le calcul des BLUP avec le modèle mixte classique, puis dans la méthode basée sur le pédigrée utilisée comme témoin lors de l'évaluation empirique de la sélection), c'est essentiellement le pédigrée connu des Deli qui a été utilisé. La seule amélioration apportée a été l'ajout d'une génération d'autofécondation pour l'individu Deli illégitime PO4953D, pour lequel toute la généalogie faisait défaut.

Par ailleurs, en ce qui concerne l'évaluation de la sélection génomique, la comparaison la plus pertinente est à faire avec la méthode actuelle, c-à-d le BLUP classique tel qu'il a été appliqué ici, utilisant les données phénotypiques des hybrides et le pédigrée connu des parents.

L'adaptation du logiciel MOLCOANC aura donc essentiellement servi ici à l'estimation des N_e réalisées de la population Deli.

La N_{eC} réalisée est une moyenne sur la période couverte par le pédigrée alors que N_{eC} calculée par le DL est valable pour la génération des parents des individus génotypés (II. J). Pour le groupe B, et en particulier pour les La Mé, le pédigrée est globalement court et il est donc normal de trouver des valeurs semblables. Le cas des Deli est considéré à la fin de la discussion de Cros et al. (2014), à l'Annexe 3.

Le ratio N_{eC} / N_{eP} est égal à 1 dans une population où la reproduction est aléatoire (Cervantes, Pastor, et al., 2011) et est supérieur lorsque la population est subdivisée. Ce ratio vaut 2.5 dans le groupe B, 1.8 ou 2.0 chez les Deli (selon le nombre de générations supposées dans le pédigrée) et 1.9 chez les La Mé. Comme à l'issue des calculs de parenté moléculaire, il ressort que la subdivision la plus forte s'observe dans le groupe B, du fait de sa structure en sous populations (La Mé et les Yangambi relativement isolés).

Compte tenu des faibles valeurs de N_e mises en évidence, des efforts devraient être faits pour réduire la consanguinité et diversifier la base génétique des populations d'amélioration, et ce afin de maintenir les perspectives de progrès génétique. Par exemple pour le groupe A, Cochard et al. (2009) ont montré que la population Deli pouvait se diviser en deux sous-populations (III. A. 3. c). Les Deli considérés ici sont des Deli Dabou et des Deli

SOCFIN et ils appartiennent à la même sous-population. Il serait bénéfique d'utiliser aussi la seconde sous-population pour développer une population d'amélioration Deli avec une N_e plus grande. Pour le groupe B, il serait utile d'avoir aussi recours à d'autres populations africaines (Cameroun, Ghana, etc.).

Comme on l'a vu précédemment (III. B. 5. b), on s'attend à ce que $16N_eL$ marqueurs permettent d'atteindre la précision maximale permise compte tenu des autres facteurs influençant la précision de la sélection génomique (taille de la population d'apprentissage, apparentement avec les candidats à la sélection, etc.). Si on considère une N_e de 5 dans les populations parentales, sachant que la longueur L du génome du palmier à huile est d'environ 2 M, le nombre de marqueurs nécessaires serait de 160. La couverture SSR disponible dans nos populations semble donc suffisante pour mener une étude de sélection génomique.

La parenté moléculaire calculée ici est basée sur les identités par état, et risque donc de surestimer les probabilités d'identité par descendance (II. D). Comme nous avons des données sur des populations considérées comme non apparentées, il serait possible d'apporter une correction selon la méthode développée par Bernardo (1993). Elle consiste à corriger l'indice de similarité entre deux individus en tenant compte des proportions moyennes de marqueurs communs entre chaque individu et ceux des populations qui ne leur sont pas apparentés. Cette approche a été améliorée par Maenhout et al. (2009) qui ont proposé un estimateur d'identité par état pondérée (WAIS, *weighted alikeness in state*), qu'il serait intéressant de tester sur nos données.

Comme on s'y attendait, le déterminisme des caractères étudiés apparait essentiellement additif, avec un ratio entre variance des ASC et variance génétique totale entre croisements du même ordre que ce qu'ont obtenu Purba et al. (2001) en analysant un ensemble de tests sur descendance comparable au dispositif considéré ici. On voit que la méthode traditionnelle basée sur des tests en croisement analysés par un modèle mixte basé sur le pédigrée permet effectivement de réaliser une sélection très précise des individus des deux groupes parentaux pour leur AGC. La faible variance des ASC indique que leur valorisation permettrait un accroissement très limité du gain génétique. Par ailleurs, l'estimation des ASC est associée à une forte imprécision et tout ceci justifie que, dans la pratique, la sélection ne porte que sur les AGC.

Le caractère essentiellement additif des composantes de la production de régimes NR et PM et leur très forte corrélation additive antagoniste indique qu'un modèle d'hétérosis sans dominance tel que décrit précédemment (II. F) est bien adapté. L'existence d'une part de dominance chez ces deux caractères, bien que faible, montre que ce modèle doit toutefois se combiner à un (faible) effet d'hétérosis qui existe chez NR et PM.

CHAPITRE V. PRECISION EMPIRIQUE DE LA SELECTION GENOMIQUE

Ce chapitre est une expérimentation destinée à estimer empiriquement la précision de la sélection génomique. Elle a fait l'objet d'une publication rédigée en anglais parue en ligne dans *Theoretical and Applied Genetics* en décembre 2014.

Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.)

David Cros, Marie Denis, Leopoldo Sánchez, Benoit Cochard, Albert Flori, Tristan Durand-Gasselin, Bruno Nouy, Alphonse Omoré, Virginie Pomiès, Virginie Riou, Edyana Suryana, Jean-Marc Bouvet

D. Cros (✉), M. Denis, B. Cochard, A. Flori, V. Pomiès, V. Riou, J. M. Bouvet
CIRAD, UMR AGAP (Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit), 34398 Montpellier, France
e-mail: david.cros@cirad.fr

L. Sánchez
INRA, UR0588, UAGPF (Forest Tree Improvement, Genetics and Physiology Research Unit), 45075 Orléans, France

E. Suryana
P.T. SOCFINDO Medan, Medan 20001, Indonesia

A. Omoré
INRAB, CRAPP, Pobè, Benin

B. Nouy, T. Dura*nd-Gasselin*
PalmElit SAS, 34980 Montferrier sur Lez, France

Key message

Genomic selection empirically appeared valuable for reciprocal recurrent selection in oil palm as it could account for family effects and Mendelian sampling terms, despite small populations and low marker density.

Abstract

Genomic selection (GS) can increase the genetic gain in plants. In perennial crops, this is expected mainly through shortened breeding cycles and increased selection intensity, which requires sufficient GS accuracy in selection candidates, despite often small training populations. Our objective was to obtain the first empirical estimate of GS accuracy in oil palm (*Elaeis guineensis*), the major world oil crop. We used two parental populations involved in conventional reciprocal recurrent selection (Deli and Group B) with 131 individuals each, genotyped with 265 SSR. We estimated within population GS accuracies when predicting breeding values of non progeny tested individuals for eight yield traits. We used three methods to sample training sets and five statistical methods to estimate genomic breeding values. The results showed that GS could account for family effects and Mendelian sampling terms in Group B but only for family effects in Deli. Presumably, this difference between populations originated from their contrasting breeding history. The GS accuracy ranged from -0.41 to 0.94 and was positively correlated with the relationship between training and test sets. Training sets optimized with the so-called CDmean criterion gave the highest accuracies, ranging from 0.49 (pulp to fruit ratio in Group B) to 0.94 (fruit weight in Group B). The statistical methods did not affect the accuracy. Finally, Group B could be preselected for progeny tests by applying GS to key yield traits, therefore increasing the selection intensity. Our results should be valuable for breeding programs with small populations, long breeding cycles or reduced effective size.

Keywords: Genomic selection, oil palm, yield, relatedness, GBLUP, Mendelian sampling term

Introduction

Genomic selection (GS) is a form of marker assisted selection that can improve breeding schemes in plants and animals. It relies on dense genome wide marker coverage to produce genomic estimated breeding values (GEBV) from a joint analysis of all markers. GEBV are obtained by summing up estimates of marker effects or through a realized additive relationship matrix markers. The model is calibrated using individuals with known phenotypes and genotypes (training set), and subsequently used to produce GEBV on a different set of selection candidates that were only genotyped (test set) (Meuwissen et al., 2001). Depending on the breeding system, genetic gain per year is expected to increase because of the higher accuracy of GS as compared to conventional selection, shorter generation intervals with the early testing of selection candidates (especially when conventional selection involves progeny testing) and/or higher selection intensity (especially when phenotyping is a limiting factor). Statistical methods to estimate GEBV use two types of information: additive genetic relationships between training and test sets and LD between markers and QTL (Habier et al., 2007, 2010). The GEBV thus implicitly take the two parts of the breeding value of an individual into account, ie the average value of its parents (family effects) and the Mendelian sampling term (within-family effects). The Mendelian sampling term originates from the random sampling of the parental gametes. It represents the deviation between the additive value of the individual and the average breeding value of its parents (Daetwyler et al., 2007, 2013). The accuracy of GS, which is the correlation between GEBV and true breeding values, is affected by linkage disequilibrium (LD) between markers and quantitative trait loci (QTL), the relationship between training and test sets, the number of individuals in the training set, the statistical method to estimate GEBV, the trait heritability and the distribution of underlying QTL effects (Lorenz et al., 2011; Grattapaglia, 2014).

Currently, few empirical studies have assessed the GS potential in plant species with long breeding cycles (>10 years) (see Grattapaglia, 2014; Isik, 2014 for reviews), and to our knowledge only Zapata-Valenzuela et al. (2012) assessed GS with a limited number of phenotyped individuals. Oil palm (*Elaeis guineensis*) is a diploid, monoecious and allogamous perennial crop with high GS potential due to its conventional breeding system. It is the major world oil crop, with a production over 55 Mt (USDA, 2014) which is expected to further increase substantially as demand for palm oil could be between 120 and 156 Mt in 2050 (Corley, 2009). Currently, oil palm genetic improvement is generally based on the reciprocal recurrent selection (RRS) scheme designed in the 1950s (Gascon and de Berchoux, 1964). It relies on two populations, the Deli (of Asian origin) and the Group B (a mixture of African populations), used as parents of the commercial hybrids. Phenotypically, these populations differ, with Deli producing a small number of large bunches and Group B a large number of small bunches. Also, they have different histories: Deli has fewer founders (4) than Group B (15 - 20) and was submitted to more generations of selection, inbreeding and genetic drift, as Deli founders were planted in 1848 and Group B founders were collected in the first half of the 20th century. In addition, the mass selection that was applied in both populations differed in its intensity, traits of interest, etc. The RRS scheme aims at increasing oil yield, which is a function of bunch number, bunch weight and fruit to bunch, pulp to fruit and oil to pulp ratios. Candidate palms sampled from full-sib families in each of the two populations are

progeny tested in Deli × Group B crosses and evaluated in extensive field trials, in order to get reliable estimated breeding values (EBV, with accuracy between 0.80 and 0.90 for all yield components). The best individuals are selected within each parental population to produce the following generation and commercial hybrids. Therefore, conventional breeding in oil palm is costly and time consuming, with a long breeding cycle (around 20 years, while sexual maturity is reached at around 3 years of age) and a limited number of tested individuals. The private oil palm breeding sector is thus seeking a practical implementation of GS that would increase the annual rate of genetic gain. In this species, the main GS challenge is currently to achieve accuracy of GEBV high enough to allow selecting among individuals that have not been progeny tested, despite the small training sets that are available (<200 progeny tested individuals per population and generation). The growing number of transcriptomic studies (eg Tranbarger et al., 2012; Tee et al., 2013; Dussert et al., 2013) and the fact that the whole genome sequence is now available (Singh et al., 2013) will facilitate the development of large numbers of SNP markers, which in turn will boost GS applications. Oil palm could therefore become a model species for GS in plants, especially for species with a long breeding cycle and/or limited phenotypic records.

The only study in which the GS potential was investigated in oil palm is a simulation by Wong and Bernardo (2008), which yielded promising results. However, their results might not be easily generalized as the simulated breeding populations resulted from selfing a hybrid between two inbred lines, while real breeding populations are more complex. Therefore, an empirical study appeared necessary.

Our objective here was to assess the potential of GS in the context of current oil palm RRS breeding by obtaining the first empirical estimate of GS accuracy using the largest EBV and genotype datasets available for the species. Specifically, we investigated a within population GS strategy for Deli and Group B populations (see Figure 23 for details). For this purpose, we used individuals with microsatellite (SSR) genotypes and deregressed EBV (DEBV) that were obtained from interpopulation progeny tests. Within each population, cross-validation was performed in order to assess the prediction accuracy of GS, as the ability to predict the breeding value of individuals that were not progeny tested. We aimed at quantifying the effects of four parameters on the GS accuracy: (1) the relationship between training and test sets: we used three methods to define the training and test sets on the basis of their genetic relationships; (2) the genetic architecture of the trait, for which we studied eight yield traits; (3) the statistical method used to estimate the GEBV: we compared five statistical methods known to behave differently depending on the genetic architecture of the traits; and (4) the population: our study included Deli and Group B populations, assuming that their contrasted history would lead to genetic differences like LD profile and genetic architecture of traits.

Materials and Methods

The data available (ie individuals with both EBV and genotypes) represented 131 Deli and 131 Group B individuals. The progeny tests to obtain EBV required around 350 ha and

15 years of data records, illustrating the difficulty to build large training sets in oil palm. Individuals were genotyped with 265 SSR.

Populations and molecular data

All individuals belonged to families from the commercial oil palm breeding program of PalmElit, a leading oil palm breeding company (www.palmelit.com). The Deli population originated from four ancestral oil palms planted in 1848 in Indonesia and was selected for yield at least from the early 20th century. Inbreeding was commonly used, by selfing or mating related selected individuals (Corley and Tinker 2003). The 131 Group B individuals included 94 La Mé (Côte d'Ivoire), 24 Yangambi (Democratic Republic of the Congo), 4 La Mé × Yangambi, 7 La Mé × Sibiti (Democratic Republic of the Congo, related to Yangambi) and 2 Nigeria individuals. The base of African populations was also formed by few founders, collected during the first half of the 20th century. In particular, the Congo population originated from around ten individuals, one of which being over 50% represented, and La Mé originated from three individuals (Cochard et al., 2009). African populations were also submitted to inbreeding and selection for yield. The inbreeding effective population size (N_e) calculated with LDNE software (Waples and Do, 2008) as described by Cros et al. (2014) was 5.0 ± 1.1 (SD) for Deli and 3.9 ± 0.8 for Group B. The 131 Deli and 131 B individuals spread over three generations. From the eldest to the most recent generation, the individuals were as follows: eight Deli and seven of Group B, 89 Deli and 99 of Group B and 34 Deli and 25 of group B (see pedigrees in Figure S1 and Figure S2). The 15 individuals of the eldest generation were selected at the end of the first RRS cycle and the others were tested in the second cycle.

The individuals were genotyped with 265 SSR (Billotte et al., 2005; Tranbarger et al., 2012). The number of polymorphic SSR markers was 220 in Deli and 260 in Group B, leading to marker densities of one SSR per 7.9 and 6.7 cM, respectively, based on a genome length of 1,743 cM (Billotte et al., 2005). The polymorphic SSR had 2.7 ± 0.8 alleles in Deli and 6.2 ± 2.2 in Group B. For GS analysis, alleles with a frequency of under 0.05 in the training set were excluded. BEAGLE 3.3.2 software (Browning and Browning, 2007) was used for imputing sporadic missing SSR genotypes, which represented 1.74% of the data in Deli and 2.90% in Group B. Molecular coancestry (ie kinship) calculated according to Lynch (1988) and Li et al. (1993) was on average 0.58 in Deli (range 0.42 - 0.96) and 0.39 in Group B (0.12 - 0.92). The heat maps of the molecular coancestry matrices are presented in Figure 24 and indicated that the populations were highly structured.

Estimation of breeding values used as data records for GS

Prior to the GS analysis, we calculated the estimated breeding values (EBV) of the 131 Deli and 131 Group B individuals. This was done using the traditional BLUP methodology (T-BLUP) (Henderson, 1975), using their pedigree and the data of their progeny tests, conducted in a large-scale experiment at Aek Loba (Sumatra). Eight traits were considered at adult age: bunch number (BN), average bunch weight (ABW), fruit to bunch (F/B), pulp to fruit (P/F), kernel to fruit (K/F) and oil to pulp (O/P) ratios, number of fruits per bunch (NF)

and the average fruit weight (FW). The details about the computation of the EBV are given in Appendix. Estimates of the narrow-sense heritability (h^2) of each trait were obtained at the experimental design level from the T-BLUP analysis as the ratio of additive variances (σ^2_{Deli} and σ^2_B for Deli and Group B, respectively) to the total phenotypic variance of crosses. The EBV accuracy was computed from the prediction error variance reported with the BLUP of each individual, the additive variances and inbreeding coefficients (See Appendix). T-BLUP shrinks individual EBV towards the parental average, thus invalidating their use as records for GS or association studies. This shrinkage, however, can be corrected by deregressing the EBV. The use of deregressed EBV (DEBV) as data records for genomic selection has proved to be beneficial compared to the use of EBV (Ostersen et al., 2011; Gao et al., 2013). Deregressed EBV can be obtained directly from existing evaluations. It appears to be equivalent to the use of other indirect methods commonly used, like daughter-yield deviation in livestock (Thomsen et al., 2001). To transform EBV into DEBV we used the approach described in Garrick et al. (2009), as previously applied in eucalyptus (Resende et al., 2012).

Definition of training and test sets

In order to investigate the range of GS accuracy that could be achieved within a given population, we used three strategies to define training and test sets: (1) K-means clustering was used to separate the individuals into five subpopulations. This method minimizes the relationships between training and test sets and maximizes the relationship within training sets (Saatchi et al., 2011). It was expected to give the lower bound in the accuracy range; (2) A within family strategy with random partition of each full-sib family into five groups, hence each individual in the test set had full-sibs in the training set. The aim was to achieve high accuracy associated with a high relationship between the training and test sets; and (3) using an optimization method, termed "CDmean" (Rincent et al., 2012), that maximized the expected accuracy of GS for the dataset. This defined a training set optimized from marker data so as to achieve the highest GS accuracy when using the remaining individuals as the test set.

In all cases, the GS model was fitted using the DEBV and genotype of the training individuals, and the fitted model was used to obtain the GEBV of the test individuals from their genotype. The K-means clustering and Within-Family strategy allowed a five-fold cross-validation. Each combination of four groups was used in turn as a training set to estimate the GEBV on individuals in the fifth group, which was used as the test set. Consequently for K-means clustering and Within-Family strategies, five GS accuracy values were obtained for each population and trait. With CDmean, only one accuracy value was obtained for each population and trait as this method yields a single optimized sample of the genotyped individuals.

The K-means clustering strategy uses a dissimilarity matrix between individuals computed from the additive relationship matrices ($\boldsymbol{A}$) of each population, according to Saatchi et al. (2011). Five clusters were made in each population using the Hartigan and Wong algorithm, implemented in the R software (R Core Team, 2014). The CDmean method (Rincent et al., 2012) optimizes sampling of the training set among the genotyped individuals. The method allocated the individuals into training or test sets based on their genotype, in a

way that maximizes the expected accuracy of GS for the dataset. The optimization criterion is the mean of the generalized coefficients of determination (CD) of contrasts between each non-phenotyped individual and the population mean. The optimization algorithm is a simple exchange algorithm. The parameters used were the additive and residual variances obtained from the mixed model that produced the initial EBV, with 16,000 iterations and 80% of the individuals assigned to the training set.

The relationship between the training and test sets was measured by the maximum additive genetic relationship between individuals in the test and training sets (a_{max}) (Saatchi et al., 2011). In order to measure the relationships between individuals in a training set, a_{max} was also calculated within training sets ($a_{max\ TRAINING}$).

Table 1 summarizes the characteristics of the obtained training sets.

Genomic selection statistical methods and control pedigree-based model

We used five GS statistical methods to obtain the GEBV of test individuals. For comparative purposes, we also used a control pedigree-based model (PBLUP) to check the usefulness of marker information. PBLUP was applied in the same way as GS statistical methods, except that PBLUP used a pedigree-based additive relationship matrix instead of marker data to model the dependencies between training and test individuals.

The GS methods were the GBLUP, which is a linear mixed model (Henderson, 1975) using a molecular additive relationship matrice $\boldsymbol{G}$ (Lynch, 1988; Li et al., 1993), and four Bayesian methods: Bayesian Lasso regression (BLR) (Park and Casella, 2008; de los Campos et al., 2009), Bayesian random regression (BRR) (Pérez et al., 2010), BayesCπ (Habier et al., 2011; de los Campos et al., 2013) and BayesDπ (Habier et al., 2011; de los Campos et al., 2013). GBLUP and BRR methods assume a common variance σ^2_m for all markers (actually alleles here, as SSR are multiallelic). BLR estimates a variance specific to each allele. In BayesCπ and BayesDπ, a priori an allele effect is zero with a probability π and non-zero either with variance common to all alleles (BayesCπ) or allele-specific variance (BayesDπ) with probability $(1-\pi)$. In both approaches, the parameter π is considered unknown and estimated from the data. As the aim of this study was to predict DEBV, we only fitted the additive effects of each allele in our models. Due to the multiallelic nature of SSR markers, the molecular data were arranged into a matrix $\boldsymbol{Z}$ with alleles in columns (instead of markers when dealing with SNP) and individuals in rows, and elements Z_{ij} = 0, 1 or 2 depending on the number of alleles j for individual i. For all GS methods, we used an heterogeneous residual variance depending on the reliability of the EBV on the individual, as described in Garrick et al. (2009).

For GBLUP, the following model was used:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Xg} + \boldsymbol{e}$$

where y is the vector of DEBV, μ is the overall mean, $\mathbf{1}$ is a column vector of 1s, $\boldsymbol{g}$ is the vector of random additive values of individuals (GEBV) following $N(0, \boldsymbol{G}\sigma^2_g)$ with σ^2_g the additive variance and $\boldsymbol{G}$ the molecular relationship matrix, $\boldsymbol{X}$ is a diagonal design matrix and $\boldsymbol{e}$ is the vector of residual effects following $N(0, \sigma^2_e)$, with σ^2_e the residual variance. $\boldsymbol{G}$ contained the similarity indices of Lynch (1988) and Li et al. (1993), which can be applied to multiallelic markers and are unbiased estimators of coancestry when assuming founder alleles

were unique (Eding and Meuwissen, 2001). This is equivalent to $\boldsymbol{G} = \boldsymbol{Z}^{t}\boldsymbol{(Z)} / 4q$, with q the number of markers and $^{t}\boldsymbol{(Z)}$ the transpose matrix of $\boldsymbol{Z}$.

The BRR, BLR, BayesCπ and BayesDπ statistical methods estimated allele effects using the following model:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zm} + \boldsymbol{e}$$

where $\boldsymbol{m}$ is the vector of allele effects. Using estimated allele effects, the GEBV of individual i was given by:

$$\hat{g}_i = \sum_{j=1}^{n} Z_{ij}\hat{m}_j$$

where n is the total number of alleles and $\hat{m}_j$ is the estimated posterior mean effect of allele j over the post burn-in iterations.

For BRR, σ^2_m and σ^2_e had scaled inverse chi-square priors with specific degrees of freedom and scales and $\boldsymbol{m}$ had a normal prior $N(0, \sigma^2_m)$. For BLR, σ^2_e followed a scaled inverse chi-square prior distribution, m_j followed a conditional Gaussian prior distribution $N(0, \tau^2_j \sigma^2_e)$ with variance specific at each allele j where τ^2_j followed an exponential prior with rate $\lambda^2 / 2$ and the regularization parameter λ^2 followed a gamma prior. For BayesCπ, π followed a beta prior, σ^2_e followed a scaled inverse chi-square prior, the conditional prior distribution of m was a Gaussian distribution $N(0, \sigma^2_m)$ with probability $(1 - \pi)$ and a null value with probability π, *and* σ^2_m followed a scaled inverse chi-square prior. For BayesDπ, π followed a beta prior, σ^2_e followed a scaled inverse chi-square prior, the conditional prior distribution of m_j was a Gaussian distribution $N(0, \sigma^2_{mj})$ with probability $(1 - \pi)$ and a null value with probability π, and the allele-specific variance σ^2_{mj} followed a scaled inverse chi-square prior with the scale parameter treated as unknown and following a Gamma(1,1) prior. For all Bayesian methods, we used 50,000 iterations with the first 12,500 as burn-in and a thinning interval of 10.

The control pedigree-based model (PBLUP) was similar to GBLUP, except that it used the $\boldsymbol{A}$ matrix of additive relationship computed from the pedigrees, instead of $\boldsymbol{G}$. As PBLUP only used pedigrees to model genetic covariances between individuals, it did not account for Mendelian sampling term, giving identical EBV to full-sibs in the test set. Thus, PBLUP only differentiated families, not individuals within families. Consequently, we expected GS to reach a higher accuracy than PBLUP by accounting for both family effects and Mendelian sampling terms. In order to check whether the GBLUP accuracy was higher than PBLUP, we carried out one-tailed paired sample t-tests for each of population-trait combination.

We used R-ASReml (Butler et al., 2009) for GBLUP and PBLUP and the BGLR R package (de los Campos et al., 2013) for BLR, BRR, BayesCπ and BayesDπ.

Prediction accuracy and bias of GEBV

Given that the true breeding values (TBV) were unknown, it was not possible to estimate the GS accuracy, which is the correlation between GEBV and TBV. Instead, we estimated the prediction accuracy, which is the correlation between GEBV and DEBV.

However, as the accuracy of EBV in oil palm progeny tests is high (between 0.80 and 0.90) the prediction accuracy was expected to be close to theoretical GS accuracy.

When investigating the correlation between the accuracy and a_{max}, a box-cox transformation was applied to *(accuracy + 1)* using $\lambda = 3$ to achieve the normality of residuals. In order to identify the factors affecting the GS accuracy, an analysis of variance (ANOVA) was performed using box-cox transformed accuracy. The factors included in the ANOVA were the GS statistical methods, the methods to define training sets, the populations, the traits, the interactions between traits and populations and the replicates (within traits and methods to define the training sets).

The prediction bias was estimated by comparing the regression of DEBV on GEBV and its expected value of one. The slope of the regression of GEBV on DEBV was thus calculated for each trait using simple linear regression.

Results

Effect of the GS statistical method on accuracy and bias of GEBV

ANOVA indicated that there was no effect of the GS statistical method on accuracy. This point is illustrated by Supplementary Figure S3, which shows almost perfect positive linear correlations between the accuracies of the five statistical methods used for genomic predictions, with Pearson correlations ranging from 0.982 to 0.995. Therefore, all the methods yielded similar accuracy regardless of the population, trait and training set definition method. The same conclusion was reached with respect to the bias, which was similar for all the statistical methods (not shown). Consequently, we only considered the results of the GBLUP method in the rest of the study.

GBLUP accuracy compared to the control pedigree-based (PBLUP) model

In Group B, GBLUP accuracy was significantly higher than that of PBLUP for three traits (ABW, BN and FW) (Figure 25). For those traits, the accuracy gain with GBLUP ranged from 22% (FW) to 89% (ABW). This superiority could be explained by the fact that GBLUP accounted for both family effects and Mendelian sampling terms (individual deviations from family effects). For the other traits, GBLUP and PBLUP accuracies were similar, indicating that markers failed to capture Mendelian sampling differences and only revealed, at best, family effects. The ability of GBLUP to capture Mendelian sampling terms was also illustrated by the existence of significant correlations between GEBV and DEBV within-full-sib families. For example, in the replicate 5 of K-means clustering in group B, the within-family GBLUP accuracy was high for ABW in the two large full-sib families that were present in the test set, reaching 0.508 in the selfing of individual LM2T (20 individuals, $P<0.05$) and 0.562 in the LM2T × LM5T cross (14 individuals, $P<0.05$). In this example, the GBLUP accuracy reached 0.588 in the whole test set and outperformed PBLUP (accuracy - 0.123). However, GBLUP was not able to estimate Mendelian sampling terms in all cases. Thus, in the replicate 3 of K-means clustering in group B, the GBLUP accuracy was null for

F/B in the largest full-sib family that was present in the test set (accuracy of 0.016 in the selfing of LM5T on 10 individuals), and GBLUP accuracy in the whole test set was not higher (0.433) than the PBLUP accuracy (0.506). In the Deli population, GBLUP failed to outperform PBLUP for all traits. Even when the mean GBLUP accuracy was higher than PBLUP (F/B, K/F, P/F), this was not significant. Therefore in Deli test individuals, the markers (like the pedigree) only allowed estimating, at best, the family effects. The Deli population was also the one presenting the lowest within-family phenotypic variance, which was on average 40% lower in Deli than in Group B, ranging from 69% lower for O/P to 14% lower for F/B (see example of ABW in Figure 26), less polymorphic markers and lower marker density; and all these conditions could impair the advantage of GBLUP over that of PBLUP.

The superiority of GBLUP over PBLUP increased when a_{max} decreased (not shown) as PBLUP could not perform well when the genetic covariances between individuals were too small (ie when a_{max} was small), while GBLUP could.

The population effect on the GBLUP accuracy was not significant. On average over all traits the GBLUP accuracy was 0.50 in Deli and 0.55 in Group B. However, the population affected the PBLUP accuracy, which was the lower in Group B (0.47) than in Deli (0.54).

Factors affecting the GBLUP accuracy

There was marked variation in the GBLUP accuracy, which ranged from negative (-0.41) to high positive values (0.94), depending on the method to define the training set, replicates, traits and traits within populations. ANOVA showed that the method to define the training set had the strongest effect on accuracy (F=155.1), followed by interactions between traits and populations (F=7.0), trait (F=5.7) and replicates (F=3.0) (P<0.001 for all factors).
The effect of the method to define the training set and replicates actually reflected the effect of the relationship between training and test sets. In all populations, CDmean gave a high maximum additive relationship between training and test set (a_{max}), the Within-Family method gave intermediate a_{max} and clustering led to low a_{max}, with one replicate with a_{max} close to zero (Figure 27). The maximum additive relationship within training sets was also affected by the method to define the training sets, but to a lesser extent than a_{max}. A significant positive correlation between the accuracy of GBLUP and a_{max} was found for almost all population-trait combinations (Figure 28). The highest accuracies were obtained when the training set was optimized with CDmean. They reached 0.79 on average, ranging from 0.49 (P/F in Group B) to 0.94 (FW in Group B). When the training set was defined by K-means clustering, the accuracy was low, at 0.29 on average, ranging from 0.04 for O/P in Deli to 0.49 for FW in Group B. For some training sets defined with clustering (in particular for those with very small a_{max} with the training individuals), negative accuracies were found. We assumed this reflected different linkage phase between marker and QTL alleles for distantly related individuals present in the training and test sets.

A significant value for the trait-population interaction in the ANOVA analysis was obtained because the O/P accuracies in Deli (0.29) was much lower than other accuracy values and because the FW accuracy in Group B was much higher (0.71) (see Figure S4 for

the complete interaction diagram). The trait effect was due to the accuracy of O/P (mean 0.42) significantly lower than the accuracy of BN (mean 0.60).

Estimates of h^2 ranged from 0.21 (O/P in Deli) to 0.57 (ABW in Group B) (Supplementary Figure S5). There was a significant positive correlation between accuracy and h^2 in Group B, although weak (P=0.020, R^2=0.62). It was not significant in Deli. This was consistent with the findings of Grattapaglia (2014), who indicated that although h^2 affected the GS accuracy, its effect was actually secondary. Moreover, we used DEBV as records and the deregression process reduces the effect of h^2 on GS accuracy (Saatchi et al., 2011).

GS bias

A strong correlation was found between accuracy and bias, indicating that the higher the accuracy, the lower the bias. GEBV was unbiased from accuracies of around 0.6 and above (Supplementary Figure S6).

Discussion

This paper presents the first experiment on genomic evaluation in a set of two oil palm breeding populations currently used in conventional reciprocal recurrent selection. We found that genomic selection (GS), in the conditions of this experiment, gave accuracies at least comparable or superior, depending on traits, to those from pedigree-based model (PBLUP) when predicting the EBV of individuals with no data records (ie not progeny tested). Superiority in accuracies was attained in one of the populations (Group B) and for some traits, due to the ability of GS to estimate the Mendelian sampling term of individuals that were not progeny-tested, as indicated by the significant correlations between GEBV and DEBV that could be observed in full-sib families. For the second population (Deli), however, results were not as conclusive, with no detectable differences between accuracies between the two evaluation methods across targeted traits. In any case, GS appeared to be a valuable method for oil palm breeding, as it opens the door to reduce the load of phenotypic evaluation and the generation interval, both important constraints in the current breeding program.

The only study to date that focused on the feasibility and potential of GS for oil palm is the simulation work of Wong and Bernardo (2008). They concluded that the genetic gain per year of GS would be higher than that of phenotypic selection if the training set had more than 50 individuals. Such a small training set was detrimental to the GEBV accuracy compared to that of conventional evaluation, but as the length of the breeding cycle with selection on markers alone was shortened to six years, the genetic gain per year ultimately increased. A novel aspect brought by our analysis is the assessment of GS in true breeding conditions, using real data from two selected populations that represent the complexity that can be found in the breeding programs for the species. We showed that reducing the need of progeny tests only to the generation used to train the GS model would be more difficult than in the forecited simulations, where training was done over the result of single crosses. Some of the critical points regarding the performance of GS highlighted by our analyses are developed in the following sections.

The range of accuracy of GS we had in our study was comparable to the values obtained by Zapata-Valenzuela et al. (2012) in loblolly pine. They also studied the implementation of GS in a perennial crop with a small number of individuals (149), using a population with a low N_e (resulting from a structured mating design). Although they had a larger number of markers (3,406 SNP), they hypothesized that their GS accuracy relied more on familial linkage than on historical LD between markers and QTL. In their case GS accuracy was similar to conventional phenotypic selection. This was not the case in our study as conventional phenotypic selection in parental populations of oil palm has a high accuracy (between 0.80 and 0.90).

Information captured by markers

We assumed that the differences in performance of GBLUP relative to PBLUP among traits and populations, as well as the effect of trait by population interactions on the GBLUP accuracy, resulted from different phenotypic variances among populations and traits and from the difference in marker informativeness among populations. These differences were likely a consequence of the contrasted history of the two populations. Each population suffered from different bottleneck events, were subjected to independent selection regimes and distinct drift effects. Compared to Group B, the Deli population had a narrower genetic base of four founders and a longer history of artificial selection, drift and inbreeding. This likely explains the fact that Deli had the lowest within-family phenotypic variance and consequently, Mendelian sampling terms are expected to be of smaller magnitude than in the Group B. As another consequence of its history, the Deli had the lowest number of alleles per marker, which was on average 2.4 compared to 3.7 in Group B, and the lowest marker density (due to more monomorphic markers than in Group B). Finally, the markers used in this study were not informative enough for the Deli population to give good estimates of the realized additive relationships and this did not allow GBLUP to generate good estimates of Mendelian sampling terms for individuals that were not progeny tested; which lead to GBLUP not performing better than PBLUP. By contrast, the Group B had higher within-family phenotypic variance and higher total number of alleles than Deli, indicating that GBLUP could have a marked advantage over PBLUP in Group B. In other words, in Group B, compared to Deli, the Mendelian sampling terms of individuals not progeny tested were easier to estimate with GS as they had a higher magnitude and because the markers were more informative.

GS utilizes the additive genetic relationship between training and test sets and LD between markers and QTL to estimate GEBV, thus accounting for both family effects and Mendelian sampling terms (Habier et al., 2007, 2010; Daetwyler et al., 2013). The proportion of GS accuracy coming from relationship and LD varies depending in particular on the marker density and training set size. Jannink et al. (2010) showed that when a small training size (400 individuals) was combined with a small number of markers (400 SNP), a large part of the GBLUP accuracy came from the relationship. This is what we observed empirically. Cochard (2008) showed that the LD was higher in the Deli than in the African populations used in this study for short distances (below 30-35 cM) and was lower for longer distances. He also found that the LD, measured by the correlation coefficient between SSR markers (r^2),

decayed to less than 0.10 within approximately 17 cM in Deli, 10 cM in La Mé and 7 cM in Yangambi. Consequently, given the marker density in our two parental populations, the LD between adjacent markers was higher in Deli than in Group B. However, as GBLUP could only estimate Mendelian sampling terms in Group B, this indicated that LD was actually not the key parameter in our dataset. LD information is of greater interest for the practical application of GS as it is more persistent than the relationship over generations (Habier et al., 2007). The challenge is thus to increase the proportion of accuracy due to LD. This could be achieved by increasing the size of the training set and marker density.

The highest superiority of GBLUP over PBLUP was obtained when a_{max} was small, ie when, according to the pedigree, the training and test sets were loosely related or unrelated. One information to bring into consideration here is the fact that pedigrees were not deep enough as to reach the base of unrelated founders (for example in Deli the pedigree did not trace back to the four founders of 1848), allowing for some individuals to appear erroneously as unrelated according to pedigree records. In such cases, marker information brought advantages to GS, as they could capture hidden relationships between individuals, as well as possible identical-by-state QTL and markers between individuals.

Surprisingly, the PBLUP accuracy could be high, in particular when optimizing the training set with CDmean. Obviously this does not mean that progeny tests are useless, but it does indicate that there was a strong genealogical structure in our breeding populations, as a consequence of inbreeding and selection. The high accuracies obtained with PBLUP were due to the ability of the pedigree to model this structure. Using GS to select among individuals that were not progeny tested, if high accuracies are obtained solely as a result of family differences, only selection between families can be carried out, with no possibility of selecting within families. This would lead to a marked increase in inbreeding and reduce future genetic progress. Therefore, in order to be useful for practical breeding, GS must account for the two parts of breeding values, ie family effects and Mendelian sampling terms. Our results stress the need for a control pedigree-based method when evaluating the potential of GS, as it helps in assessing the ability of GS to account for Mendelian sampling terms.

We studied eight traits, assuming there should be variations in genetic architecture among them, in particular in the number of QTL, as some traits could be less complex than others. Several authors using real data reported that there was no effect of the statistical method used to estimate GEBV (Heslot et al., 2012; Kumar et al., 2012; Daetwyler et al., 2013). This could be due to the limited number of training individuals and markers, or could have resulted from the fact that the true genetic architecture actually involved large numbers of QTL for all traits.

Definition of training sets

Using K-means clustering, within-family and CDmean to define the training and test sets gave more valuable information on the GS accuracy than simple replicates with random assignment, as the different methods substantially affected the relationship between the training and test sets. We observed a marked decrease in GS accuracy with decreasing maximum additive relationships (a_{max}) between the training and test sets. This was similar to

the results obtained by Habier et al. (2010) in Holstein cattle with large training sets (2,096 and 1,408) and a large number of SNP (54,001).

The use of the optimization algorithm, based on a CDmean maximized relationship between training and test sets and a minimized relationship within the training set, yielded the highest GS accuracies. CDmean therefore appeared to be the best method. In a practical use of GS, all individuals in the generation(s) used to calibrate the model would be genotyped at juvenile stage and CDmean would be applied to identify the subset of individuals to progeny test. This subset would make an optimized training population, ie the one maximizing the GS accuracy. Finally, selection would be made based on GEBV among all individuals, either both genotyped and progeny-tested or only genotyped. In our study, we defined an optimized training set specific to each trait using the corresponding heritability (h^2) values. Obviously, for practical application, it would be necessary to use a mean value of h^2 over traits that must be selected. This should have a negligible effect on the accuracy, as Rincent et al. (2012) showed that the CDmean method is robust to h^2 variation, which we also observed here as the training sets were very similar among traits.

Practical prospects for oil palm

In the perspective of an optimal use of GS that would allow making selection on markers alone and limiting the use of progeny tests to the training of the GS model, oil palm breeding should evolve toward a reciprocal recurrent genomic selection breeding scheme integrating marker data to increase the selection intensity and decrease the length of breeding cycles (Figure 23). In this scheme, GS could be applied among individuals that have not been progeny tested and that belong to the same generation as the training individuals or to the following generation(s). Using selection candidates highly related to the training set (for instance full-sibs) would correspond to the situation we studied with the Within-Family and CDmean strategies, which proved to be favorable in terms of accuracy. However, if selection candidates are loosely related to the training set (although from the same population), our results with the K-means strategy indicated that GS would fail, with accuracy very low, and possibly negative. This case could occur for example when companies exchange breeding material after several generations of independent selection. As less effort would be required for genotyping candidate individuals than progeny testing them, GS could increase the selection intensity as compared to conventional breeding. In addition, if the GS accuracy is high enough to conduct selection solely on markers in the generation(s) following training, the length of the breeding cycle would decrease, as progeny tests would only be made in the generation used to train the model. However, this would only be possible if the GS accuracy were high enough for all the yield components. In Group B, the accuracy for some key oil yield components (especially average bunch weight [ABW] and bunch number [BN]) in the test sets was higher with GS models than with the pedigree-based control model (PBLUP). The markers could thus be used for preselection before progeny tests by identifying genetically superior individuals for ABW and BN, which would subsequently be progeny tested to finalize selection on these two traits (as the accuracy of EBV from conventional progeny tests is higher than the GEBV accuracy), and for phenotypic-based selection on the other yield components with lower GBLUP accuracy. This would increase the intensity of

selection on ABW and BN, thus increasing the rate of genetic gain for yield. Obviously, this would not tap the full potential of GS, which could only be achieved if GS reduced the need for progeny tests. This will not be possible as far as there is not a clear-cut advantage of the GS models over pedigree-based models for all yield traits. Considering that the new scheme would alternate one generation of progeny tests to calibrate the GS model with one generation of selection on markers alone, the length of two cycles would be only 60% of the current length. This new breeding scheme will be a credible alternative when, for all yield components, GS will be able to account for the Mendelian sampling terms and will have a mean accuracy over two cycles higher than 60% of the highest accuracy obtained currently in reciprocal recurrent selection, ie higher than 0.54.

In order to validate our new breeding scheme integrating GS, the first points to investigate are the effects on accuracy of larger training sets and a larger number of markers, to identify how many individuals and markers are required for GS to outperform pedigree information for all traits and populations. Larger training sets could be achieved by adding each new generation of progeny tested individuals to the existing training set. The increase in the number of markers could be achieved by genotyping all individuals with next generation sequencing or with a SNP chip, which could be developed using the whole genome sequence now available (Singh et al., 2013). Another crucial question to be addressed is the decrease in GS accuracy when applying the model in the subsequent generations following training. Moreover, our study used data that were collected in a single environment, which likely led to an upward accuracy bias due to a common error component in both GEBV and EBV (Lorenz et al., 2011). The first results of progeny tests of the next breeding cycle will be available within a few years. They will be used to estimate the effect of a larger training set and a larger number of markers on the GS accuracy, as well as the decrease in accuracy when applying GS models in a test set generated by the crossing of individuals selected in the training generation.

To our knowledge, this is the first empirical study of GS with SSR markers. In the near future we will rather use SNP markers, as this will make the analysis easier (due to the biallelic nature of SNP), decrease the cost per data point and allow faster genotyping. In a simulation, Solberg et al. (2008) concluded that two to three times more SNP were required to achieve the same accuracy as with SSR. In our oil palm breeding populations, the difference in the number of SNP and SSR necessary to reach a given accuracy will surely be smaller, as the polymorphism of SSR was low, with some markers actually being biallelic.

We used a two-step approach, first obtaining deregressed estimates of the additive value (DEBV) of the progeny-tested Deli and Group B parents and, second, using these values as data records in the GS model to measure the GS accuracy when predicting the DEBV of individuals not progeny tested. An alternative would have been to implement a single step methodology, using the whole dataset (ie the phenotypic data of the progenies) and, for a given training set of parents, considering only the crosses made with these parents (ie discarding from the analysis the data of the progenies of the test parents) to directly predict the GEBV of the test parents. Although such an approach was appealing, it could not be implemented here. Indeed, the available data represented only one experimental design, and this had to be analyzed as a whole. Analyzing just one part of the experimental design would have lead to a highly unbalanced dataset, with parents and trials becoming disconnected from

the rest of the experimental designs and biases appearing in the estimates of the non genetic effects. In real life situations, oil palm breeders would use the results of a whole experimental design to calibrate the GS model, therefore taking advantage of its qualities (connectedness between trials and between parents, balance in the number of crosses per parent, etc). We chose a two-step procedure in order to mimic such a situation. Obviously, when data will become available from several experimental designs, we will likely adopt a single step approach.

Authors' contribution

DC carried out data analysis and wrote the paper, with the contribution of MD, LS and JMB. BC and VP carried out genotyping work. TDG, ES, BC, AO, BN, AF and VR made field experiments and data collection.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

We acknowledge SOCFINDO (Indonesia) and CRAPP (Benin) for planning and carrying out the field trials with CIRAD (France) and authorizing use of the phenotypic data for this study. This research was partly funded by a grant from PalmElit SAS. We thank P. Sampers, C. Carrasco-Lacombe, A. Manez and S. Tisné for help in genotyping, L. Dedieu for reviewing the manuscript as well as two anonymous reviewers and C.C. Schön for helpful comments.

APPENDIX. Estimation of parental breeding values

The mating design of the progeny tests consisted of 445 Deli × Group B crosses made according to an incomplete factorial design. The crosses were evaluated in 26 trials planted between 1995 and 2000. The experimental designs of the trials were RCBD with five or six blocks and balanced lattices of rank four or five. The bunch production was measured on 30,872 palms and bunch quality on 21,525 palms. Eight traits were studied. The bunch number (BN) and average bunch weight (ABW) were measured every ten days on palms from ages 6 to 11. The annual cumulative BN and mean annual ABW were used in analysis. The median number of progenies with bunch production data was 169 per Deli parent (ranging from 25 to 743) and 141 (23-859) per Group B parent. The fruit to bunch (F/B), pulp to fruit (P/F), kernel to fruit (K/F) and oil to pulp (O/P) ratios, the number of fruits per bunch (NF)

and the average fruit weight (FW) were measured on two bunches at ages five and six on a sample of at least 24 palms per cross. The median number of bunches analyzed was 327 per Deli parent (ranging from 69 to 1,358) and 309 per Group B parent (73-1,149).

EBV were computed as traditional pedigree-based BLUP (T-BLUP) predictors of the random effects $\boldsymbol{a_A}$ and $\boldsymbol{a_B}$, using a mixed model of the form:

$$\boldsymbol{y = X\beta + Z_1 a_{Deli} + Z_2 a_B + Z_3 b + Z_4 c + Z_5 p + Z_6 k + e}$$

where $\boldsymbol{y}$ is the vector of data records for the trait being analyzed, β the vector of fixed effects (general mean, trial and block within trial), $\boldsymbol{a_{Deli}}$ and $\boldsymbol{a_B}$ vectors of general combining ability of Deli ~ $N(0,\ 0.5\boldsymbol{A_{Deli}}\ \sigma^2_{Deli})$ and Group B individuals ~$N(0,\ 0.5\boldsymbol{A_B}\ \sigma^2_B)$, respectively, $\boldsymbol{b}$ the vector of the incomplete block within block and trial effects ~$N(0,\ \boldsymbol{I}\sigma^2_b)$, $\boldsymbol{c}$ the vector of specific combining ability of single crosses ~$N(0,\ \boldsymbol{D}\sigma^2_c)$, $\boldsymbol{p}$ the vector of permanent environmental effects used to take repeated measures into account ~$N(0,\ \boldsymbol{I}\sigma^2_p)$, $\boldsymbol{k}$ the vector of elementary plot effects ~$N(0,\ \boldsymbol{I}\sigma^2_k)$ and $\boldsymbol{e}$ the vector of residual effects ~$N(0,\ \boldsymbol{I}\sigma^2_e)$. $\boldsymbol{X}$, $\boldsymbol{Z_1} - \boldsymbol{Z_6}$ are incidence matrices. $\boldsymbol{A_{Deli}}$ and $\boldsymbol{A_B}$ are matrices of additive relationships among Deli and Group B individuals, respectively, computed from pedigrees. $\boldsymbol{D}$ is the matrix of dominance relationships among crosses computed from the pedigree, with value between crosses Deli × B and Deli' × B' equal to $f_{Deli,\,Deli'} \times f_{B,B'}$, where $f_{Deli,\,Deli'}$ and $f_{B,B'}$ are the coefficient of coancestry between the Deli and Group B parents. $\boldsymbol{I}$ is an identity matrix. For BN and ABW, the model also included a fixed age effect and a random age within cross effect $\boldsymbol{a}$ ~$N(0,\ \boldsymbol{D} \otimes \boldsymbol{I}\sigma^2_a)$. This model was based on the model of Stuber and Cockerham (1966) for hybrids between unrelated populations, as previously used in oil palm by Purba et al. (2001). The R-ASReml package (Butler et al., 2009) for R (R Core Team, 2014) was used to obtain variance component estimates and EBV of all individuals.

The accuracy of the general combining ability a_i of an individual i (actually a_{Deli_i} or a_{B_i} depending on the population of origin of i) is given by $r_{a,\hat{a}_i} = \sqrt{1 - \frac{PEV_{a_i}}{0.5(1+F_i)\ \sigma^2_a}}$, where PEV_{ai} is the prediction error variance associated with a_i, $0.5(1+F_i)$ is the diagonal of the relationship matrix used in the mixed model (ie $0.5\boldsymbol{A_{Deli}}$ or $0.5\boldsymbol{A_B}$, depending on the population of origin of i), F_i is the inbreeding coefficient and σ^2_a is the additive variance (ie σ^2_{Deli} or σ^2_B, depending on the population). This formula was used to compute the mean accuracy of the general combining ability of the 131 Deli and 131 Group B parents used in the GS analysis, which was 0.89, ranging from 0.83 ± 0.06 (SD) for O/P in Deli to 0.93 ± 0.04 for K/F in Group B.

CHAPITRE VI. GAIN GENETIQUE SUR LE LONG TERME DE LA SELECTION GENOMIQUE

Ce chapitre est une simulation destinée à estimer le gain génétique annuel de différentes formes de sélection génomique sur quatre générations, par rapport à la sélection récurrente actuelle. Elle a fait l'objet d'une publication en anglais soumise le 14/12/2014 et qui est en cours de relecture.

Long-term genomic selection for heterosis without dominance in multiplicative traits: case study of bunch production in oil palm

David Cros, Marie Denis, Jean-Marc Bouvet, Leopoldo Sánchez.

D. Cros (✉), M. Denis, J.M. Bouvet
Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit (UMR AGAP), CIRAD, 34398 Montpellier, France
e-mail: david.cros@cirad.fr
Tel.: +33-467615800

L. Sánchez
Forest Tree Improvement, Genetics and Physiology Research Unit (UR AGPF), INRA, 45075 Orleans, France

Key message

Simulations based on oil palm showed that reciprocal recurrent genomic selection could outperform phenotypic reciprocal recurrent selection for hybrid performance resulting from multiplicative interactions between additive and antagonistic components.

Abstract

To study the potential of genomic selection for heterosis resulting from multiplicative interactions between additive and antagonistic components, we focused on oil palm, where bunch production is the product of bunch weight and bunch number. We simulated two realistic breeding populations and compared over four generations the current reciprocal recurrent selection (RRS) with reciprocal recurrent genomic selection (RRGS). All breeding strategies aimed at selecting the best individuals in parental populations to increase bunch production in hybrids. For RRGS, we obtained the parental genomic estimated breeding values using GBLUP with hybrid phenotypes as data records and population specific allele models. We studied the effects of four RRGS parameters on selection response and genetic parameters: (1) the molecular data used to calibrate the GS model: in RRGS_PAR, we used parental genotypes and in RRGS_HYB we also used hybrid genotypes; (2) frequency of progeny-tests (model calibration); (3) number of candidates and (4) number of genotyped hybrids in RRGS_HYB. We concluded that RRGS could increase the annual response to selection compared to RRS by decreasing the generation interval and by increasing the selection intensity. With 1,700 genotyped hybrids, calibration every four generations and 300 candidates per generation and population, response to selection of RRGS_HYB was 71.8% higher than RRS. RRGS_PAR with calibration every two generations and 300 candidates was a relevant alternative, as a good compromise between annual response, risk around expected response, increase in inbreeding and cost. Finally, RRGS required inbreeding management because of higher annual increase in inbreeding than RRS.

Keywords: genomic selection, GBLUP, oil palm, simulation, hybrid, reciprocal recurrent selection

Introduction

Genomic selection (GS) (Meuwissen et al., 2001) is the state-of-the-art method of marker assisted selection for complex traits. GS relies on dense genome wide marker coverage to produce genomic estimated breeding values (GEBV) from a joint analysis of all markers. GEBV can be obtained using a realized additive relationship matrix computed from markers, in what is called the GBLUP method (VanRaden, 2007; Habier et al., 2007). The GS model is calibrated using individuals with known phenotypes and genotypes (training set) and predicts the GEBV of selection candidates. For phenotypically evaluated candidates, the interest of GS lies in its ability to give GEBV with higher accuracy ($r_{\hat{A}A}$, the correlation between true and estimated breeding values) than the EBV traditionally obtained through expected additive relationships computed from the pedigree. GS also gives GEBV of selection candidates that were only genotyped, allowing selection without phenotypic evaluation. This reduces the length of the generation interval (L), especially if conventional breeding requires long progeny-tests, and increases the selection intensity (i) when the cost of phenotyping is higher than the cost of genotyping. Consequently, the annual response to selection, which is given by $r_{\hat{A}A} \times i \times \sigma_a / L$ (with σ_a the additive standard deviation), can be higher with GS than with phenotypic selection.

GS can be used to increase the performance of interpopulation plant hybrids and crossbred animals. Kinghorn et al. (2010) simulated a crossbreeding system where GS was applied to select among parental lines in order to increase heterosis in crossbred animals for a trait with dominance at QTL. The highest response to selection was obtained in their study with reciprocal recurrent genomic selection (RRGS), which consisted in using phenotypes and gametotypes of crossbred individuals to estimate line specific marker effects. Heterosis in a complex trait can also result from the multiplicative interaction between additive and negatively correlated components (Schnell et Cockerham, 1992; Gallais, 2009 p68-71). For example, this can be the case for yield in crops as a product of fruit weight and fruit number, or plant height as a product of number and length of internodes. In such a case, dominance at QTL is not necessary to explain the heterosis in the multiplicative trait. As GS proved to be efficient for single additive traits in many studies, it could also be beneficial in the case of multiplicative interactions between complementary parental components. However, this potential benefit over conventional phenotypic selection has not been quantified so far. Oil palm is an interesting model for this purpose. In oil palm, the bunch production is the product of bunch weight (BW) and bunch number (BN), two negatively correlated and mostly additive traits (Gascon et al., 1966; Corley et Tinker, 2003). Oil palm breeding currently relies on reciprocal recurrent selection (RRS) between two heterotic populations showing complementary characteristics for BW and BN. One of them is the Deli population of Asian provenance and the other is African, commonly the La Mé population (Côte d'Ivoire). Deli palms have small number of large bunches, while La Mé have large number of small bunches. This results in heterosis in the hybrids for bunch production, which is at least 25% higher than in the parental populations (Gascon et de Berchoux, 1964).

The potential of GS in oil palm has been evaluated in two studies. One of these (Cros et al. (2014)) is an empirical work where the accuracy of GS is estimated through a cross-validation approach. They implemented population specific GS models with the progeny-

tested parents as training-sets, using their deregressed EBV and genotypes. This approach required little genotyping effort, as the number of progeny-tested parents was reduced (<200 per population and generation). On the other hand, these small numbers lead to small training sets, consequently with a detrimental effect for GS accuracy. The small number of individuals available to train the GS model is a problem common to many species having costly and time-consuming field trial evaluations, in particular in perennial crops (for instance around 0.5 ha per cross and 15 years of data record in oil palm). An alternative to enlarge the training set could be to include also hybrid individuals, to take advantage of the allelic segregation existing within hybrid crosses due to heterozygosity in parents. For this purpose, we can implement a GS analysis taking into account hybrid gametotypes and parental genotypes, like in Kinghorn et al. (2010). Furthermore, Cros et al. (2014) applied a cross-validation approach in a single generation, while it would be more interesting to assess the potential of GS over the long term, taking into account not only the accuracy of selection but also the generation interval and selection intensity. For this purpose, computer simulation is useful, in particular for a species with long breeding cycles and extensive field trials (Sun et al., 2011).

The second study on the potential of GS in oil palm is a simulation work (Wong et Bernardo, 2008). They evaluated by simulation the potential of GS in oil palm over three generations and concluded that this method gave higher annual response to selection than phenotypic and marker assisted selection. However, they made simplifying assumptions: considering a single additive trait instead of the multiple trait approach of actual programs, and a parental population that resulted from the selfing of a hybrid between inbred lines, which did not correspond to existing oil palm breeding populations. Also, Wong and Bernardo (2008) only studied the response to selection, without considering the evolution of genetic parameters. Therefore, new studies on the potential of GS in oil palm are still justified, notably to cover more complex and general situations, and over several generations.

The aim of our study was to quantify the potential of RRGS as an alternative to conventional RRS, when the objective is to improve hybrid performance resulting from the multiplicative interaction between additive antagonistic components. For this purpose, we focused on the example of the oil palm species, for which current conventional breeding is characterized by long generation interval, due to progeny-tests, and by small populations. We simulated two realistic (complex) oil palm breeding populations with complementary characteristics for bunch production components (bunch weight [BW] and bunch number [BN]). We used these populations to compare several RRGS breeding schemes to traditional RRS over four generations, with the aim of improving the hybrid performance of interpopulation crosses for bunch production. More precisely, the simulations investigated the effects of four GS parameters on the response to selection obtained with GBLUP in terms of bunch production: (1) the molecular data and associated GS model: in RRGS_PAR, we used parental genotypes to compute $\boldsymbol{G}$ matrices specific to each parental population and a model predicting two independent random effects of general combining abilities, one for each parental population, and in RRGS_HYB we used genotypes of both parents and hybrid individuals to compute one $\boldsymbol{G}$ matrice taking into account the parental origin of marker alleles and a model predicting a single random effect of breeding values; (2) the frequency of progeny-tests (*ie* calibration of GS model): every generation, every two generations or every four generations; (3) the number of candidate individuals: 120 or 300; and (4) for

RRGS_HYB, the number of genotyped hybrids: 300, 1,000 or 1,700. We also studied the evolution of the genetic parameters in the parental populations: selection accuracy and additive variance for BW and BN, their genetic correlation and inbreeding.

Materials and Methods

Simulation overview

The overall simulation process is summarized in Figure 29. It involved three steps: (i) the simulation of an equilibrium base population, (ii) the simulation of initial breeding populations derived from this base population and having realistic genetic characteristics compared to current real oil palm breeding populations, and (iii) the simulation of the breeding strategies (reciprocal recurrent selection [RRS] and two strategies of reciprocal recurrent genomic selection [RRGS_PAR and RRGS_HYB]) applied to the initial breeding populations for four generations. The simulated genome had a length of 17 M and 16 chromosomes. The mutation rate was 10^{-5} per bp per meiosis, and was kept constant for the rest of the simulation.

The simulations were made with R software version 3.0.2 (R Core Team, 2014) and the HaploSim package (Coster et Bastiaansen, 2010). The following paragraphs explain in details the three steps of the simulation.

Simulation of equilibrium base population

We simulated a population over 2,400 discrete generations with a constant size of 200 individuals having equal contribution to the following generation and reproducing by random mating with the exclusion of selfing. In the first generation, 20,000 bi-allelic loci (SNP) with equal distances between adjacent loci and equifrequent 0 and 1 alleles per locus were simulated across the genomes. The functions in HaploSim allowed simulating meiosis between two parental haplotypes in order to produce a new individual. Mutation-drift equilibrium was assessed in the final generation with the distribution of allelic frequencies, expected to follow a typical U-shaped curve. In the last generation there were 17,572 segregating SNP. A single base population was generated this way and used as a starting point for the rest of the simulation process.

In the last generation, segregating SNP with minor allele frequency (MAF) above 0.1 were chosen at random to be causative mutations (QTL). We assumed that the negative genetic correlation existing between BW and BN resulted from pleiotropy and consequently some QTL were randomly chosen to have pleiotropic effects. To study the effect of QTL number (n_{QTL}) and percentage of pleiotropic QTL (p_{QTL}) on the results, we simulated n_{QTL} = 100, 500 and 1,000 QTL per trait and p_{QTL} = 60%, 75% and 90%. We assumed simple additive architecture for BW and BN. For pleiotropic QTL, the QTL substitution effects for BW (α_{BW}) and BN (α_{BN}) were drawn from a normal bivariate distribution. This distribution was defined assuming a correlation of -0.9 and QTL variance equal to $\sigma^2_{a(BW)BP} / n_{QTL}$ for BN and to $\sigma^2_{a(BN)BP} / n_{QTL}$ for BW, with base population variances $\sigma^2_{a(BW)BP}$ = 6 and $\sigma^2_{a(BN)BP}$ = 12

chosen by trial and error so that $\sigma^2_{a(BW)}$ and $\sigma^2_{a(BN)}$ in the simulated initial breeding populations matched with the actual values. For non-pleiotropic QTL, α_{BN} and α_{BW} were drawn from normal distributions using the same QTL variances as for pleiotropic QTL. The breeding (additive) value for each individual and the additive variance were defined according to the quantitative genetic model of Falconer and Mackay (1996). As we considered bi-allelic QTL, the breeding value for a trait at a QTL in a given population was equal to $-2p\alpha$ for homozygous genotype 00, $(1-2p)\alpha$ for heterozygote 01 and $2(1-p)\alpha$ for homozygote 11, where p was the frequency of allele 1 in the population and α the substitution effect of the QTL for the corresponding trait. The breeding value of each individual was obtained by summing the breeding value at each QTL across all QTL. In a given population, the intrapopulation additive variances were $\sigma^2_{a(BW)} = \sum_{QTL(BW)} 2p(1-p)\alpha_{BW}^2$ and $\sigma^2_{a(BN)} = \sum_{QTL(BN)} 2p(1-p)\alpha_{BN}^2$. As we assumed purely additive genetic determinism, they were equal to twice the interpopulation additive variances. The residual variances $\sigma^2_{e(BN)}$ and $\sigma^2_{e(BW)}$ were calculated with the formula $\sigma^2_{e(BW)} = \sigma^2_{a(BW)}(1-h^2_{BP}) / h^2_{BP}$ and $\sigma^2_{e(BN)} = \sigma^2_{a(BN)}(1-h^2_{BP}) / h^2_{BP}$ using a base population heritability $h^2_{BP} = 0.8$ for each trait, chosen so that h^2 in the simulated initial breeding populations matched with the actual h^2. Environmental effect on BW and BN were generated from normal distributions with mean zero and variances $\sigma^2_{e(BW)}$ and $\sigma^2_{e(BN)}$. We assumed that $\sigma^2_{e(BW)}$ and $\sigma^2_{e(BN)}$ were the same for the two populations and they were kept constant for the simulation. The residual correlation between BW and BN was assumed to be zero. The phenotype was assumed to be the sum of the breeding value, environmental effects and mean value of the population. Initially, the mean value of the population for traits BW (kg) and BN were set to 15, in order to avoid negative values for bunch production and to obtain realistic phenotypic values for BW and BN in the simulated initial breeding populations.

Simulation of initial breeding populations

The simulation process adopted to create the initial breeding populations from the equilibrium base population tried to mimic what is known of the history of the Deli and La Mé populations (Corley et Tinker, 2003; Cochard, 2008). We proceeded by trial and error to set up the parameters of the simulation that were not know from the literature or the real data (for example the selection intensity for the mass selection), in order to generate breeding populations with genetic parameters that were consistent with actual parameters.

The equilibrium base population was randomly divided into two populations A and B of 100 individuals each. For 100 generations, A and B populations evolved independently and kept constant size. Mating was at random without selfing. Each population had a different selection regime so that they had a divergent evolution for the two traits: increasing BW in A and increasing BN in B. The parents of the individuals in a given generation were sampled in the previous generation, where the probability for each individual to be chosen as a parent was proportional to its phenotypic value.

After these first 100 generations, four individuals were taken at random in population A to simulate the bottleneck event at the origin of the actual Deli population, which originated from four oil palms planted in Indonesia in 1848. This was followed by three generations of random mating without selfing with 25 individuals per generation, and by six generations

with increasing number of individuals (50, 50, 60, 75, 100 and 150 individuals per generation). During these last six generations, mass selection was applied on bunch production. For mass selection, bunch production of each individual was computed as the product between BN and BW phenotypes. The best 70% individuals were selected and randomly mated with the exclusion of selfings to produce the following generation. Similarly in population B, after the first 100 generations of divergence 19 individuals were taken at random to simulate the bottleneck event at the origin of the actual La Mé population in Côte d'Ivoire in the 1920s. This was followed by two generations with increasing number of individuals (75, 150) and mass selection on bunch production. Mass selection was implemented in the same way as for Deli, but selecting the top 30% individuals.

At that point, the simulated Deli and La Mé populations were submitted to two generations of RRS for bunch production, simulating what occurred in the real oil palm breeding populations from the 50s. The principle of RRS is to select among candidate individuals based on their EBV, obtained from progeny tests. Here, EBV were simulated as values correlated to the true breeding values, with a correlation of 0.8 in first RRS cycle and 0.9 in second RRS cycle, corresponding to the accuracy of actual oil palm progeny-tests. We selected in each parental population the top 20 individuals giving the crosses with highest expected bunch production. The expected bunch production of each cross between the progeny-tested Deli and La Mé was calculated as the product between the mean parental EBV for BW and BN. In each population, selected individuals reproduced by random mating with selfings allowed according to a diallel design in which 80% of crosses were made. In the last generation (generation 0 in Figure 29), 300 individuals were produced per population, uniformly distributed among crosses. They will hereafter be referred to as initial breeding populations. Genotypes at SNP and QTL from the initial breeding populations, as well as pedigree information of the last four generations in Deli and last two in La Mé were retained to be used in the final step of the simulation (comparison of RRS and RRGS).

The simulation process was repeated several times from the allocation of QTL to the generation of initial breeding populations. Runs were kept only if the genetic parameters in the simulated initial breeding populations were close to the real values of the current Deli and La Mé breeding populations. This calibration was made on: the fixation index F_{st} between Deli and La Mé, profiles of linkage disequilibrium (LD), narrow-sense heritabilities (h^2), additive variances for BW and BN, and genetic correlation between BW and BN. The Table 2 summarizes the observed values and the mean values and standard deviations (SD) obtained for the different genetic parameters in the replicates kept for the study (five replicates per combination of n_{QTL} and p_{QTL}). The real values of F_{st}, interpopulation additive variances and genetic correlation were obtained from the dataset described in Cros et al. (2014). It consisted in 131 Deli crossed with individuals from various African populations, including 94 La Mé; in order to progeny test them at Aek Loba (Sumatra). They were genotyped with 265 SSR markers. Weir and Cockerham estimate of F_{st} was computed using hierfstat R package (Goudet, 2013) in the simulated data and with diveRsity R package (Keenan et al., 2013) in the real data. Although the simulated populations included SNP markers, the real value found with SSR markers could be used to calibrate the simulations, as the SSR had polymorphisms close to those of SNP. The actual genetic correlation between BN and BW and additive variances for BN and BW in parental populations were computed from the hybrids phenotypic

values using a mixed model analysis. For LD the absolute values are affected by the marker type, so we only used the profile of LD curves to compare simulated and real populations. As references we used the LD curves calculated by Cochard (2008), which showed higher LD in Deli over short distances (below 30-35 cM) and higher LD in La Mé for longer distances. For h^2, we used as target values the mean h^2 for BN and BW in Deli and La Mé reported in the literature (Hardon et Thomas, 1968; Thomas et al., 1969; Meunier et al., 1970; Ooi et al., 1973). Also, the simulation runs where a single QTL explained over 20% of the total additive variance were discarded, as this was considered unrealistically high. Finally, five replicates were produced for each combination of number of QTL (n_{QTL}) and percentage of pleiotropic QTL (p_{QTL}).

Simulation of reciprocal recurrent selection and reciprocal recurrent genomic selection

The initial Deli and La Mé breeding populations were used as starting points to compare conventional RRS with RRGS over four generations, in terms of genetic gain for bunch production in hybrid individuals (response to selection) and evolution of genetic parameters in the parental populations (selection accuracy and additive variance for BN and BW, genetic correlation between BN and BW, inbreeding).

In reciprocal recurrent selection (RRS), the EBV of the Deli and La Mé selection candidates were obtained from the analysis of their progeny-tests. At each generation the progeny-tests involved 120 Deli and 120 La Mé. The mating design for progeny-tests was an incomplete factorial design with 300 crosses (*ie* 2.5 crosses per parent, with variation in number of crosses per parent made as small as possible). We simulated 45 individuals per cross, with their resulting SNP genotypes, QTL genotypes and breeding and phenotypic values for BW and BN. The progeny-tests were analyzed with a bivariate mixed-model to obtain the EBV of the 120 Deli and 120 La Mé for BW and BN. The model was of the form:

$$\begin{bmatrix} y_{BW} \\ y_{BN} \end{bmatrix} = \begin{bmatrix} \mu_{BW} \\ \mu_{BN} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_{Deli(BW)}} & 0 \\ 0 & \boldsymbol{Z_{Deli(BN)}} \end{bmatrix} \begin{bmatrix} a_{Deli(BW)} \\ a_{Deli(BN)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_{La\ Mé(BW)}} & 0 \\ 0 & \boldsymbol{Z_{La\ Mé(BN)}} \end{bmatrix} \begin{bmatrix} a_{La\ Mé(BW)} \\ a_{La\ Mé(BN)} \end{bmatrix} + \begin{bmatrix} e_{BW} \\ e_{BN} \end{bmatrix}$$

with y_{BW} and y_{BN} vectors of phenotypic values of the 13,500 hybrid individuals for BW and BN, μ_{BW} and μ_{BN} overall means of hybrid individuals for BW and BN, e_{BW} and e_{BN} vectors of residual effects for BW [~ $N(0, \boldsymbol{I}\sigma^2_{e(BW)})$] and BN [~ $N(0, \boldsymbol{I}\sigma^2_{e(BN)})$]. The vectors of general combining ability of BW and BN in Deli $a_{Deli(BW)}$ and $a_{Deli(BN)}$ followed a bivariate normal distribution $N(0, \begin{pmatrix} \sigma^2_{Deli_{(BW)}} & \sigma_{Deli_{(BN,BW)}} \\ \sigma_{Deli_{(BN,BW)}} & \sigma^2_{Deli_{(BN)}} \end{pmatrix} \otimes 0.5\boldsymbol{A_{Deli}})$, with $\sigma_{Deli(BN,BW)}$ the additive covariance between BN and BW. The vectors of general combining ability of La Mé $a_{La\ Mé(BW)}$ and $a_{La\ Mé(BN)}$ followed a similar distribution with population specific parameters $\boldsymbol{A_{La\ Mé}}$, $\sigma^2_{La\ Mé(BW)}$, $\sigma^2_{La\ Mé(BN)}$ and $\sigma_{La\ Mé(BN,\ BW)}$. $\boldsymbol{A_{Deli}}$ and $\boldsymbol{A_{La\ Mé}}$ were matrices of additive relationships among Deli and La Mé individuals computed from pedigrees. $\boldsymbol{Z_{Deli(BW)}}$, $\boldsymbol{Z_{Deli(BN)}}$, $\boldsymbol{Z_{La\ Mé(BW)}}$ and $\boldsymbol{Z_{La\ Mé(BN)}}$ were incidence matrices and $\boldsymbol{I}$ an identity matrix. The R-ASReml package (Butler et al., 2009) was used to obtain variance component estimates and EBV of Deli and La Mé individuals. In each population the best 20 individuals giving the crosses with highest expected bunch production were selected based on their EBV, as described for the

previous step of the simulation, and they reproduced by random mating with selfings allowed according to a half diallel design in which 80% of crosses were made (consequently, 168 different within-population crosses could be made). The number of crosses was the same for all individuals. 120 progenies per population were produced. The generation interval for RRS was 20 years.

Reciprocal recurrent genomic selection (RRGS) gave genomic estimated breeding values (GEBV) of the Deli and La Mé selection candidates, from an analysis combining their genotype and their progeny-tests or from their sole genotype. Like in RRS, the progeny-tests involved 120 Deli and 120 La Mé. The GEBV were obtained using 2,500 SNP markers with MAF>4% and the GBLUP statistical method. In the simulations we studied the effects of reducing the generation interval and increasing the selection intensity on RRGS performance. First, the reduction of the generation interval was obtained when applying RRGS in the generation(s) following the progeny-tested individuals. In this case, the selection candidates were not progeny-tested but only genotyped; and they were selected based on their sole genotype and reproduced as soon as they were sexually mature. We considered the generation interval was consequently reduced to six years (instead of 20 years in the generations where progeny-tests were made, like in RRS). To assess the potential of RRGS when used to reduce the generation interval, we varied the frequency of progeny-tests. They were simulated every generation, leading to a total number of 80 years to complete the four cycles, every two generations (52 years to complete four cycles) or every four generations (38 years to complete four cycles). The GEBV of the 120 Deli and 120 La Mé individuals in the generations with progeny-tests were predicted from the phenotypic data of hybrid individuals and molecular data of either only Deli and La Mé individuals (RRGS_PAR, see model below) or Deli and La Mé individuals plus hybrid individuals (RRGS_HYB, see model below). The GEBV of the Deli and La Mé individuals in the generations without progeny-tests were predicted in the same way, except that the phenotypic data used to calibrate the GS model were those from the last generation of progeny-tests. Second, to study the effect of increasing the selection intensity, we also applied RRGS to a number of selection candidates (300 per population) larger than the number of individuals progeny-tested (120 per population). As 168 different within-population crosses could be made, the 300 individuals were obtained by simulating one or two individuals per possible cross, up to a total of 300. In the generations with progeny-tests and when using 300 candidates, 120 individuals were randomly chosen to be progeny-tested among the 300 and the selection was made among them and their 180 non progeny-tested sibs.

In RRGS_PAR, the progeny-tests were analyzed with the same bivariate model as for RRS, except that matrices of additive relationships $\boldsymbol{A}_{Deli}$ and $\boldsymbol{A}_{La\ Mé}$ were replaced by molecular relationship matrices $\boldsymbol{G}_{Deli}$ and $\boldsymbol{G}_{La\ Mé}$ computed from parental genotypes, using observed allele frequencies (VanRaden, 2007; Habier et al., 2007) and normalized to have average diagonal coefficients equal to one (Forni et al., 2011). Therefore, RRGS_PAR used two population specific molecular relationship matrices.

In RRGS_HYB, the molecular relationship matrix $\boldsymbol{G}$ of the genotyped individuals (all the Deli and La Mé and the $n_{genotyped}$ hybrid individuals) was computed using Deli and La Mé genotypes and hybrid gametotypes, taking into account the parental origin of marker alleles for hybrid individuals. For this purpose, each SNP was converted into a multiallelic marker

with alleles 0_{Deli}, 1_{Deli}, $0_{La\ Mé}$ and $1_{La\ Mé}$. From these molecular data, $\boldsymbol{G}$ was computed according to VanRaden (2007) and Habier et al. (2007), using observed allele frequencies and a modification implemented by Legarra (pers. comm..) for the multiallelic case, and it was normalized to have average diagonal coefficients equal to one (Forni et al., 2011). The number of genotyped hybrid individuals was $n_{genotyped}$ = 300, 1,000 and 1,700. The corresponding breeding strategies were termed RRGS_HYB300, RRGS_HYB1000 and RRGS_HYB1700. The genotyped hybrid individuals were randomly sampled among the 13,500 existing hybrids, taking the same number of individuals per cross (*ie* one when $n_{genotyped}$ = 300). As progeny-tests included 13,500 phenotyped hybrids, the non-genotyped hybrids were also included in the model, as their phenotypic values contributed to estimate the GEBV of their parents. For this purpose, we used the single-step approach of Legarra et al. (2009), which implied combining the $\boldsymbol{G}$ matrix to the genealogical additive relationship matrix $\boldsymbol{A_{all}}$ of all the Deli and La Mé individuals and all the hybrids, according to: $\boldsymbol{H}^{-1} = \boldsymbol{A_{all}}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{G}^{-1} - \boldsymbol{A_{22}}^{-1} \end{pmatrix}$, with $\boldsymbol{A_{22}}$ the matrix of genealogical additive relationship matrix of the genotyped individuals (parents and hybrids). The bivariate model for RRGS_HYB was:

$$\begin{bmatrix} y_{BW} \\ y_{BN} \end{bmatrix} = \begin{bmatrix} \mu_{BW} \\ \mu_{BN} \end{bmatrix} + \begin{bmatrix} \boldsymbol{Z_{Deli(BW)}} & 0 \\ 0 & \boldsymbol{Z_{Deli(BN)}} \end{bmatrix} \begin{bmatrix} a_{BW} \\ a_{BN} \end{bmatrix} + \begin{bmatrix} e_{BW} \\ e_{BN} \end{bmatrix}.$$

The vectors a_{BN} and a_{BW} of breeding values of all individuals (hybrids and their parents) for BN and for BW followed $N(0, \begin{pmatrix} {\sigma^2}_{a_{(BW)}} & \sigma_{a_{(BN,BW)}} \\ \sigma_{a_{(BN,BW)}} & {\sigma^2}_{a_{(BN)}} \end{pmatrix} \otimes \boldsymbol{H})$, with ${\sigma^2}_{a(BN)}$ and ${\sigma^2}_{a(BW)}$ the additive variances and $\sigma_{a(BN,\ BW)}$ the additive covariance between BN and BW.

Analysis of results

We distinguished between two types of factors: technical factors under the control of the breeder (breeding strategy [RRS, RRGS_PAR and RRGS_HYB], number of selection candidates, frequency of progeny-tests and number of genotyped hybrids); and the biological factors defining the genetic architecture of the traits under selection (number of QTL, percentage of pleiotropic QTL). The effects of biological factors and the interaction between biological and technical factors are crucial because, in true situations, the actual genetic architecture is unknown. The breeder has to design a breeding program where technical factors will give the highest annual selection response, regardless of the unknown actual genetic architecture of the traits.

We called breeding scheme a combination of breeding strategy (RRS, RRGS_HYB and RRGS_PAR), frequency of progeny-tests (every generation, every two or every four generations), number of selection candidates (120 or 300) and number of genotyped hybrid individuals (0, 300, 1,000 and 1,700). Each of these combinations had 5 replicates, which had different simulated initial breeding populations.

At the end of the simulation (*ie* in generation 4 of Figure 29), we measured the cumulative response to selection in hybrid individuals, expressed in percentage of hybrid production in initial generation (generation 0), and the annual response to selection in hybrid individuals, which was the cumulative response divided by the number of years required to

carry out the four breeding generations. Analysis of variance (ANOVA) were made to study the effect of the different technical and biological factors and their interactions on the response to selection, as well as on the genetic parameters in the parental populations (selection accuracy and additive variance for BW and BN, genetic correlation between BW and BN, inbreeding).

Results

Number of genotyped hybrids in RRGS_HYB

In order to simplify the interpretation of the results, we first focused on the number of genotyped hybrid individuals, as it had a major effect on the annual response to selection in RRGS_HYB. The Table 3 presents the ranking of the breeding schemes according to their annual selection response, and showed that the annual response of RRGS_HYB increased with the number of genotyped hybrids. RRGS_HYB could outperform RRS when 1,000 or 1,700 hybrids were genotyped. The difference between RRGS_HYB1700 and RGGS_HYB1000 was generally small but was always at the advantage of RRGS1700 across all combinations. RRGS_HYB with 300 genotyped hybrids was worse than the other two alternatives with 1000 and 1700, as its annual response was always lower or equal to RRGS_PAR and RRS. Therefore, for the rest of the study we only considered the results of RRGS_HYB1700.

Accuracy of selection

The accuracy of selection with RRS was very high and remained constant over generations, around 0.967 ± 0.003 (SD), with a negligible effect of trait and population (not shown).

For RRGS, we first considered its simplest implementation *ie* when calibrating the GS model every generation and using sets of candidate individuals limited to the 120 progeny-tested individuals. In this case, the accuracy of RRGS_PAR (0.968 ± 0.008) was similar to that of RRS, while that of RRGS_HYB1700 was slightly but significantly less good ($P<0.001$), with accuracy of 0.934 ± 0.008 (see Figure 30 with the example of BN in Deli). Second, we evaluated how the accuracy of selection was affected by the absence of progeny-tests. The accuracy of the progeny-tested individuals was much higher than the accuracy of individuals not progeny-tested of the following(s) generation(s), which felt to 0.748 ± 0.058 for RRGS_HYB and even lower for RRGS_PAR, at 0.615 ± 0.101. When three generations of selection where made without calibration of the GS model, the accuracy of selection kept decreasing and this occurred at higher pace for RRGS_PAR than for RRGS_HYB (Figure 30). The accuracy of the progeny-tested individuals was also higher than the accuracy of their 180 not progeny-tested sibs, which was 0.852 ± 0.019 for RRGS_HYB and 0.744 ± 0.020 for RRGS_PAR. This can be seen when comparing levels of equivalent combinations between Figure 30A and B: the inclusion of 180 non-progeny tested individuals in the latter

dropped the general level of accuracy. For all non-progeny tested individuals, the accuracy was significantly lower with RRGS_PAR than with RRGS_HYB ($P<0.001$).

Therefore, the key results here were that the accuracy of selection was reduced in individuals that were not progeny-tested; and that the two RRGS methods behaved differently in terms of accuracy of selection for progeny-tested individuals (RRGS_PAR slightly more accurate than RRGS_HYB) and non progeny-tested individuals (RRGS_HYB much more accurate than RRGS_PAR).

Additive variance

The additive variance decreased with generations, which is a well-known effect of selection and genetic drift. The decrease in additive variance with RRS was identical to RRGS_PAR using 120 candidates and calibrated every generations (not shown). The Figure 31 shows the results obtained with RRGS in Deli when using 120 and 300 candidates, with the example of BN (similar trends were obtained for both traits and for both populations). The major factor affecting the cumulative decrease in additive variance was the number of candidates, with 300 resulting in a more rapid decrease in variance than with 120 candidates (see Figure 31A versus Figure 31B, where the additive variance decreased after four generations by 27% with 120 candidates but by 35% with 300 candidates, $P<0.001$). A similar result was observed for both populations and traits. On average, using 120 candidates decreased the additive variance in RRGS strategies by 32% while using 300 individuals lead to a decrease of 39.6%. This occurred as the number of selected individuals was kept constant, and consequently increasing the number of candidates lead to the selection of individuals with a higher average value but a lower genetic variability. We also noted that with 300 candidates the decrease in additive variance with RRGS_HYB after four generations (-41.3%) was similar than with RRGS_PAR (-37.8%), the difference being not significant. This indicated that the number of candidates was the major factor affecting the decrease in variation.

The number of QTL that were assumed in the model also had a role in the decrease of additive variance. In the same example (BN in Deli), the additive variance decreased by around 27% with either 1,000 or 500 QTL but by 37% with 100 QTL ($P<0.001$). This occurred as the simulation was designed to have the same additive variances in generation 0 regardless of the number of QTL. Consequently, when the QTL were fewer they also had stronger effects and received correspondingly higher selection pressures up to their fixation. Therefore, the fewer the number of QTL the stronger the effect of selection was on the depletion of the additive variance.

Response to selection

All the biological factors (number of QTL and percentage of pleiotropic QTL) had significant effects on the selection response at $P<0.001$. The percentage of pleiotropic QTL was the most important factor of the study. The number of QTL also had a strong effect. The selection response increased when the percentage of pleiotropic QTL decreased and when the number of QTL increased. The cumulative response was 14.7%, 18.6% and 22% with 90%,

75% and 60% pleiotropic QTL, respectively (P<0.001 for all differences), and the annual response was 0.26%, 0.33% and 0.40% (P<0.001). This resulted as the potential of genetic progress lied mostly in the fixation of the favorable allele at QTL controlling either BN or BW, rather than in the pleiotropic QTL as they had antagonistic effects on both traits. When the percentage of pleiotropic QTL decreased, the number of QTL controlling only BN or BW increased, therefore giving a higher potential of genetic progress. Similarly, the potential of genetic progress was higher when the total number of QTL controlling each trait increased.

The major technical factors affecting the selection response was the frequency of progeny-tests, followed by the number of candidates, the breeding strategy and, to a lesser extent, the interaction between breeding strategy and frequency of progeny-tests (P<0.001). Their effects are detailed in the following paragraphs.

We first compared the response to selection of RRS with RRGS strategies that differed with RRS only by the use of relationship matrices computed with markers instead of the pedigrees, *ie* considering only RRGS_HYB1700 and RRGS_PAR with calibration every generation and sets of candidates limited to 120 progeny-tested individuals. In this case, the response to selection was similar for RRGS_HYB1700, RRGS_PAR and RRS (20.5% over four generations, or 0.26% per year, see Table 3). Therefore, obtaining a higher selection response with RRGS compared to RRS could not be achieved without modifying the breeding scheme in order to reduce the generation interval or to increase the selection intensity.

Second, to study the effect of reducing the generation interval in RRGS (by the decrease in the frequency of progeny-tests), we considered RRGS_HYB1700 and RRGS_PAR with a calibration of the GS model every two or every four generations and with the same number of candidates as in RRS (120). In this case, decreasing the frequency of progeny-tests lead to a lower cumulative selection response: it was 13.9%, 17.4% and 20.2% with progeny-tests every four generations, every two generations and every generation, respectively (all differences being significant at P<0.001). With generations without progeny-tests, the accuracy of selection was reduced and consequently the cumulative response to selection decreased. The effect was opposite in annual response: it was 0.37%, 0.33% and 0.25% with progeny-tests every four generations, every two generations and every generation, respectively (P<0.001). This was due to the favorable effect of the decrease of frequency of progeny-tests on the ratio between accuracy of selection and generation interval ($r_{A\hat{A}} / L$), where the decrease in generation interval more than balanced out the decrease in accuracy while the other factors (selection intensity, additive variance) remained unaltered. Indeed, progeny-tests every two generations decreased the generation interval by 35% compared to progeny-tests every generation and decreased the accuracy of selection to 0.90, which was 4% lower than with progeny-tests each generation and 120 candidates for RRGS_HYB and 10% lower for RRGS_PAR. With progeny-tests every four generations the generation interval decreased by 52.5% and the accuracy of selection dropped to 0.84 for RRGS_HYB (10% decrease compared with progeny-tests each generation and 120 candidates) and to 0.80 for RRGS_PAR (17% decrease). Therefore, the less frequent the calibration of the GS models, the lower the cumulative response to selection but the higher the ratio $r_{A\hat{A}} / L$ and the annual response to selection. With the decrease in the frequency of progeny-tests, the relative potential of RRGS_HYB and RRGS_PAR varied, due to their different accuracy in generations with and without progeny-tests. With calibration every generation, as the

accuracy of RRGS_HYB was lower than the accuracy of RRGS_PAR, the annual response to selection of RRGS_HYB was lower than that of RRGS_PAR, although not significantly (Table 3). With progeny-tests every two generations, as the accuracy of RRGS_PAR decreased more than the accuracy of RRGS_HYB, both methods reached the same $r_{A\hat{A}} / L$ ratio and the same annual response to selection, which was significantly higher than the annual response to selection of RRS. With progeny-tests every four generations, RRGS_HYB finally outperformed RRGS_PAR in accuracy and reached significantly higher annual response to selection than the other breeding schemes (+50% compared to RRS for RRGS_HYB, +30% for RRGS_PAR).

Third, we studied for RRGS gain the effect of an increase in selection intensity, which was obtained by increasing the number of candidates, as the number of selected individuals was constant. With 120 candidates, the best 16.7% individuals were selected, while with 300 candidates only the top 6.7% were selected, which is a 2.5 fold more stringent selection intensity. This significantly increased the response to selection, as a consequence of the effect of the number of candidates on the product $r_{\hat{A}A} \times i \times \sigma_a$, with the increase in i (×2.5) being much higher than the joint decrease in $r_{\hat{A}A}$ (-8.8% in RRGS_HYB, -23.1% in RRGS_PAR) and in σ_a (-3.3%). Again, due to the superiority of RRGS_HYB over RRGS_PAR to maintain the selection accuracy for individuals not progeny-tested, the increase in the number of candidates benefited more to RRGS_HYB than to RRGS_PAR (Table 3).The annual response was always higher with 300 candidates than with 120 candidates, but this difference was only significant for RRGS_HYB (+12.8%, P<0.001 for RRGS_HYB, +7.1% for RRGS_PAR). We also found that the number of candidates significantly interacted with the percentage of pleiotropic QTL and the number of QTL on the selection response, although these interactions had less effects than the factors previously mentioned. This was not surprising as, with the largest numbers of non pleiotropic QTL per trait (either due to a high number of QTL or to a low percentage of pleiotropic QTL), 120 selection candidates were not enough to capture all the existing additive variation. In this case, using 300 candidates lead to a higher response to selection. By contrast, with a smaller number of non pleiotropic QTL, 120 candidates were enough to capture all the additive variation and an increase in the number of candidates did not increase much the selection response.

Finally, when RRGS was used for both decreasing the generation interval and increasing the selection intensity compared to RRS, the best breeding scheme was RRGS_HYB1700 with progeny-tests every four generations and 300 candidates, with a annual response to selection of 0.45% per year *ie* 71.8% higher than RRS and significantly higher than all other breeding schemes at P<0.001 (Table 3).

This advantage of the RRGS_HYB1700 scheme did not come without risks, and the higher gains presented also a larger variation of response compared to the less performing alternatives (Figure 32). The coefficient of variation (CV) for the annual response to selection of that best scheme reached 0.27, which was 35.6% higher than that of RRS, and 19.1% higher than the average CV over all the breeding schemes. The following three breeding schemes in the ranking of gain had similar levels of performance and CV for the annual response: RRGS_HYB1700 with progeny-tests every four generations and 120 candidates (annual response of 0.39% ± 0.23, +47.7% compared to RRS), and RRGS_PAR with 300

candidates and calibration every four or every two generations, which gave the same results (annual response of 0.38% ± 0.22, +45.3% compared to RRS).

Inbreeding

As expected, the inbreeding increased with the generation turn-over (see Figure 33, with the example of RRGS in Deli population). The annual increase in inbreeding (ΔF_y) with RRS was 0.41% in Deli and 0.64% in La Mé (expressed in percent of the inbreeding in the initial parental populations) (Figure 34). The factors affecting cumulative ΔF (ΔF_c) over four generations and ΔF_y were the same in both populations. ΔF_y was mostly affected by the frequency of progeny-tests, the number of candidates and the breeding strategy ($P<0.001$). A decrease in frequency of progeny-tests reduced the number of years per cycle of selection, and therefore could strongly inflate ΔF_y. The number of candidates affected both ΔF_c and ΔF_y. In Deli, ΔF_y reached 0.77% with 300 candidates against 0.64% with 120, and in La Mé it reached 1.16% with 300 candidates against 1.0% with 120 (all differences significant at $P<0.001$). Increasing selection intensity by increasing the number of candidates from 120 to 300 resulted therefore in a subsequent increase in ΔF_y, due mainly to an increase in the co-selection of related candidates. With 120 candidates, they all belonged to different full-sib families, due to the method used to mate the selected individuals. However, the sets of 300 candidates were mostly made of pairs of full-sibs, which increased the probability of having full-sib individuals among the selected ones. In addition, we noticed that RRGS_HYB was associated with slightly higher ΔF_c and ΔF_y than RRGS_PAR. In Deli, RRGS_HYB led to a ΔF_y of 0.75% against 0.69% with RRGS_PAR ($P<0.001$). In La Mé, ΔF_y was also higher with RRGS_HYB (1.12%) than with RRGS_PAR (1.10%) but this was not significant.

Finally, the four best breeding schemes identified previously in terms of annual response to selection also had high ΔF_y, with the exception of RRGS_PAR with progeny-tests every two generations and 300 candidates, thanks to its higher frequency of progeny tests.

Genetic correlation between BW and BN

The evolution of the genetic correlation between BN and BW was similar for the Deli and La Mé populations. The magnitude of the genetic correlation between BW and BN increased strongly in the generations were progeny-tests were made (Figure 35), while it usually decreased in the generations without progeny-tests. In absolute value, the increase in the generations with progeny-tests was stronger than the decrease in the generations without progeny-tests, leading to an increasing trend in the strength of the correlation over the four generations, except in the case where progeny-tests were only made in the first generation.

Discussion

We showed that reciprocal recurrent genomic selection (RRGS) was a valuable method to increase over the long term the performance for a trait showing heterosis due to the multiplicative interaction between additive and negatively correlated components.

In our oil palm case study, RRGS was superior to traditional RRS as it allowed selecting accurately individuals without progeny-tests. It led to a significant increase in the annual response to selection, through the implementation of generations of selection on markers alone and, to a lesser extent, through an increase in the selection intensity. This advantage of RRGS over RRS was a consequence of the trade-offs between selection intensity (i), generation interval (L) and accuracy of selection ($r_{A\hat{A}}$). RRGS could increase substantially the ratio $r_{A\hat{A}} \times i / L$, with the best breeding scheme (when considering only annual response to selection) being RRGS_HYB1700 with calibration every four generations and 300 candidates per generation and population. In our oil palm example, the annual response to selection of this RRGS best strategy reached 0.45%, against 0.26% in RRS. Interestingly, both RRGS strategies with and without genotyping hybrids outperformed RRS. In RRGS_HYB, the increase in annual response was strong when the number of genotyped individuals increased from 300 to 1,000, but small from 1,000 to 1,700 (although significant). Therefore, it seemed that genotyping more individuals would have been useless here. However, the number of hybrids to genotype should be under the control of the heterozygosity in the parental populations. Indeed, RRGS_HYB could perform better than RRGS_PAR because it exploited the within crosses phenotypic variation, by associating it with the within cross segregation of marker alleles. Likely, with a higher level of heterozygosity in the parental population, the phenotypic variation within crosses would increase, making more relevant the genotyping of hybrids in order to capture this variation. On the contrary, in the extreme case of fully inbred and/or related parents, genotyping hybrids would become useless due to the absence of valuable within crosses variation.

Choosing an optimal breeding scheme also requires taking into account other aspects than the sole expected annual response to selection. The RRGS_HYB1700 breeding scheme reached the highest annual response to selection but this happened at the cost of the highest variability of annual response (indicating a higher risk regarding the true genetic progress that could be achieved), and also the highest increase in inbreeding per year. Furthermore, the cost efficiency and the operational complexity must be considered. RRGS_HYB would be more difficult and more costly to implement than RRGS_PAR, because it would need the collection of more samples than RRGS_PAR, more genotyping and would require inferring gametotypes of the gametes contributed to each hybrid individual in order to identify the parental population of origin of marker alleles at each locus (Kinghorn et al., 2010), which was assumed known in the simulation. Therefore, other breeding schemes could offer interesting alternatives, as good compromises between costs, operational complexity, expected annual response to selection, risk around this expectation and evolution of inbreeding. Here, the most interesting alternative was RRGS_PAR with 300 candidates and progeny tests every two generations. Indeed, it was among the best four breeding schemes in terms of annual response to selection (0.38%), but at the same time it had a lower risk around this expected response and an increase of inbreeding that was intermediate among all the scenarios studied, together

with small cost and less operational complexity associated with RRGS_PAR strategy. Furthermore, this RRGS_PAR scenario gives the opportunity to gather the data of two-progeny tests when making the calibration in the third generation, which would likely increase the accuracy in the two last generations and therefore the mean annual selection response over the four cycles.

The relative importance of the decrease in generation interval and the increase in selection intensity depends on the characteristics of the species. In oil palm, the length of the generation interval (20 years) is mostly due to the progeny-tests, while sexual maturity is reached relatively early (from three to four years). This makes the species an excellent candidate for the implementation of early genomic evaluation, with high potential of RRGS compared to RRS. By contrast, oil palm breeding populations have rather narrow genetic basis, with effective sizes lower than 10 (Cros, Denis, et al., 2014), and this created relatively small additive variances, therefore limiting the interest of increasing the number of candidates.

Our results confirmed the usefulness of GS for oil palm, like in the simulation of Wong and Bernardo (2008). However, we extended their results to a more general situation that is closer to actual oil palm breeding programs, by applying GS to complex breeding populations and by considering two antagonistic traits BN and ABW, which is crucial in oil palm breeding. Our results were consistent with the empirical study of Cros et al. (2014). They used an approach similar to our RRGS_PAR method and the accuracies they obtained when applying GS to full-sibs of the training individuals could be compared with the accuracies we obtained here on the 180 full-sibs of the 120 individuals that were progeny-tested. They obtained mean accuracies of 0.74 for BN and BW using 105 Deli individuals to calibrate the GS model, which were the same values as in our simulation. For the La Mé population, they obtained accuracies of 0.60 for BN and 0.65 for ABW with 74 individuals to calibrate the model (unpublished results), which in this case was smaller than our accuracies (0.75 for BN and 0.74 for BW). Likely, this occurred because their training population was smaller than ours (120 here). The consistency between their empirical results and our simulations suggested that the actual genetic architecture for BN and BW could be close to the average scenario of our simulations, *ie* 500 QTL and 75% of pleiotropic QTL.

Surprisingly, RRGS_HYB with 300 hybrid individuals genotyped had poor results, with a selection response lower than RRGS_PAR. We expected that RRGS_HYB would outperform RRGS_PAR even with a small number of genotyped hybrids, due to the extra molecular information provided. We hypothesized the low performance of RRGS_HYB300 happened because with such a small number of genotyped hybrids, the molecular relationships between the genotyped hybrids and their Deli and La Mé parents (or among genotyped hybrids) were biased and not compatible with the genealogical relationships involving their non genotyped sibs. This could be due to the fact that the molecular relationships in RRGS_HYB were gametic relationships, where only one half of the molecular data of hybrid individuals were used to relate them to each of their two parents. This could possibly be resolved by using more markers or by improving the computation of the $\boldsymbol{H}$ matrix used in RRGS_HYB, which requires further investigation.

Actually, we chose different GS models for RRGS_HYB and RRGS_PAR for computational reasons: in the model chosen for RRGS_PAR, two $\boldsymbol{G}$ matrices were required.

When no molecular data from hybrids were used, the ***G*** matrices were small (from 120 × 120 to 420 × 420 here). However, using the same model for RRGS_HYB would mean having two large ***H*** matrices (from 13,740 × 13,740 to 14,340 × 14,340). This would result in extensive computation time difficult to manage in the context of a simulation study, due to the many replicates, and memory problems. For this reason, we used a different GS model for RRGS_HYB, requiring a single relationship matrice. As a consequence, in RRGS_HYB the breeding values of Deli and La Mé parents and hybrid individuals were assumed to belong to the same distribution with a common additive variance, which did not fit with the reality. However, this should not be a problem here as the mixed models were used to predict GEBV, not to estimate genetic parameters.

Management of genetic variability

We found that the best breeding schemes in terms of annual response were also associated with high annual increase in inbreeding (ΔF_y). Actually, the best breeding schemes had the highest ΔF_y, with the exception of RRGS_PAR with 300 candidates and progeny-tests every two generations that had an intermediate ΔF_y. The fact that our genomic evaluation led to higher inbreeding rates than with phenotypic evaluation is somehow in contradiction with previous results obtained in animal breeding, reviewed in Bouquet and Juga (2013), stating that genomic selection can reduce ΔF per generation compared to traditional breeding. According to them, this happened because the Mendelian sampling terms (*ie* individual genetic effects) were more accurately estimated with GS than with phenotypic selection, thus reducing the probability of selecting sibs and consequently the ΔF per generation. Depending on the reduction in the generation interval allowed by GS (which is dependent on the species), they noted this would lead to a lower or similar ΔF than traditional breeding. However, in our study the ΔF per generation was higher than with RRS. This first resulted from the fact that in oil palm RRGS allowed a very strong reduction in the generation interval (up to only 38 years to complete four cycles when calibration was made only in the first generation, compared to 80 years for traditional RRS). We assumed this was also related to the drop in accuracy observed for non progeny-tested individuals, that occurred as the calibration of the GS model was based on the progeny-tests of only 120 individuals per population. Consequently, our estimates of Mendelian sampling terms were likely not as accurate as what was obtained in animal species. However, it was not clear why RRGS_HYB lead to a higher increase in ΔF than RRGS_PAR, while it was more accurate.

Good management of the genetic variability is necessary to avoid inbreeding depression in parental populations, which has been reported in oil palm (Hardon, 1970; Corley et Tinker, 2003; Luyindula et al., 2005), and to maintain the potential of genetic progress over the long term. Furthermore, the negatively correlated BN and BW studied in this simulation are key traits for oil palm breeding and when dealing with antagonistic traits breeders must find a compromise between response to selection, variance of response induced by antagonistic traits and ΔF (Sánchez et al., 2008). Therefore, the RRGS breeding schemes we presented here should be combined with methods to manage explicitly genetic diversity and inbreeding. The most simple method of inbreeding management is to increase the number of selected individuals, which would slow down the increase in inbreeding, possibly with only

a small reduction in response to selection (Bouquet et Juga, 2013). Another option that does not lead necessarily to loses in gain is optimal contribution selection (Meuwissen, 1997) and its extension in the context of GS (Sonesson et al., 2012). This uses the genetic value of individuals and their relationships with other individuals to determine their contribution to the following generation, in order to maximize genetic gain at a desired rate of inbreeding under the assumption of random mating among selections. A step further is mate selection (Toro et Perez-Enciso, 1990; Sánchez et al., 1999), where optimum contribution is applied to mates among all candidates, so that selection and mating are simultaneously handled for improved management of inbreeding beyond what is expected by random mating. Mate selection optimizes the number of parents to be selected, the actual matings between them and the distribution of the contribution in descendants of these mates, in order to maximize the expected response to selection in the following generation while respecting a restriction on the expected increase in inbreeding.

Genomic selection model

Here we studied a GS approach to select individuals within two parental populations for their crossbred performance, like in several animal studies (Ibánez-Escriche et al., 2009; Toosi et al., 2010; Kinghorn et al., 2010; Zeng et al., 2013) and in maize (Technow et al., 2012). We used models with population-specific effects of SNP alleles, either by using a parental model with two independent parental effects (RRGS_PAR) or by distinguishing alleles depending on their population of origin (RRGS_HYB). However, Ibánez-Escriche et al. (2009) and Toosi et al. (2010) suggested that GS models using crossbred populations to predict GEBV of parental pure breeds may not need to fit breed-specific effects of SNP, especially with high marker density. However, we did not considered this point, as in Ibánez-Escriche et al. (2009) breed-specific allele models performed better than models with allele effects common to all breeds when breeds where distantly related, which was the case in our simulations.

In this study we considered that heterosis in bunch production was a consequence of the multiplicative interaction between the negatively correlated bunch number and bunch weight, both assumed here to have complete additive genetic determinism. This multiplicative interaction between complementary component differences in the parents is a model of heterosis without dominance, but heterosis in a multiplicative trait can also be due to the multiplicative interaction of component dominance (Schnell et Cockerham, 1992). In this case, dominance in the component traits generates heterosis in the complex trait, to a greater extent than the components dominance, due to the multiplicative nature of the complex trait. Here, we did not study the effect of this type of genetic determinism (or a combination of the two types). This would require further investigation, which could be done by modifying the script used for our simulations and including dominance effects in the GS models (see for instance Su et al. (2012) for a GBLUP model including dominance effects).

Genetic correlation between BW and BN

Wu and Sánchez (2011) showed that in a model associating pleiotropic and non pleiotropic QTL, the simultaneous selection on the two traits increased the magnitude of the genetic correlation. Presumably, their result applied in the case where selection was highly accurate, such as when based on progeny-tests, which was not the case here when selection was made on markers alone using a GS model calibrated with a small training set.

Authors' contribution

DC carried out simulations and analysis and wrote the paper, with the contribution of LS, MD and JMB.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We acknowledge the INRA of Orléans and Maxime Mercière (CIRAD) for granting access to their server. This research was partly funded by a grant from PalmElit SAS.

CHAPITRE VII. DISCUSSION GENERALE ET PERSPECTIVES

VII. A. Discussion générale

Les données expérimentales et simulées indiquent que la sélection génomique (SG) devrait permettre de sélectionner uniquement sur leur génotype les individus ayant les plus fortes aptitudes à la combinaison hybride. Ceci amènerait une diminution de l'intervalle moyen de génération et un accroissement de l'intensité de sélection, aboutissant à un gain génétique annuel qui pourrait dépasser de 50% celui de la méthode de sélection traditionnelle.

L'étude réalisée à partir des données réelles a montré que la précision de la SG était fortement liée à l'apparentement entre la population d'apprentissage et les candidats à la sélection. De bonnes précisions (>0.60) pourraient être obtenues sur des candidats fortement apparentés à la population d'apprentissage, tels que des plein-frères ou des descendants. Pour que cette sélection soit efficace, il faut qu'elle puisse s'appliquer au sein de familles de plein-frères, c-à-d que les GEBV des candidats tiennent compte de la valeur moyenne de leur famille et de leur terme de ségrégation mendélienne, et ce pour toutes les composantes du rendement et dans les deux groupes parentaux. Ceci n'était le cas que pour certains caractères du groupe B. Dans ces conditions, les sélectionneurs ne peuvent pas se passer de tests en croisements. Cependant, cette étude était contrainte par le fait que des données étaient disponibles sur une seule génération de tests en croisement. Par conséquent, afin d'avoir un jeu de validation sur lequel était calculé la précision de la SG, seulement 80% des individus testés en croisement étaient inclus dans le jeu d'apprentissage, soit environ 105 individus. Dans les conditions réelles, tous les individus d'une génération de tests en croisements seraient inclus dans la calibration du modèle de SG. Par ailleurs, pour concevoir une première population d'apprentissage pour démarrer la SG, il serait possible d'utiliser les données de deux générations successives de tests en croisement. Concrètement, pour une application dans les années à venir, cela reviendrait à utiliser les parents testés en croisements à Aek Loba (c-à-d les individus utilisés dans le Chapitre V) et ceux en cours de test à Aek Kwasan 2 (Indonésie), ce qui représente au total 214 Deli et 197 individus du groupe B. En 2014, ces individus ont été génotypés par GBS (*genotyping by sequencing*, voir VII. B. 3) et des données ont été obtenues pour plusieurs milliers de SNP. Des populations d'apprentissage de cette taille devraient permettre d'atteindre des précisions de sélection largement plus élevées que ce qui a été présenté au Chapitre V.

Par ailleurs, les simulations ont montré que le gain génétique annuel de la SG pourrait dépasser de 50% celui de la SRR classique, en rendant possible la suppression des tests en croisement à certaines générations. Ceci diffère donc des résultats obtenus avec les données réelles, mais cette différence peut s'expliquer par la taille plus grande des populations d'apprentissage (environ +15%), l'utilisation des données moléculaires d'hybrides, le plus

grand nombre de marqueurs et la couverture plus régulière du génome dans les simulations. On doit cependant garder à l'esprit que les simulations portaient uniquement sur la production de régimes, en négligeant leur teneur en huile. Or, sur certaines composantes de la teneur en huile des régimes, la précision de la SG pour les individus non testés en croisement pourrait être plus faible à cause de la moins grande précision des valeurs phénotypiques qui seraient utilisées pour la calibration du modèle. Ce point mérite donc d'être étudié. De plus, bien que les simulations aient été conçues pour être aussi réalistes que possibles, en respectant l'histoire connue des populations d'amélioration et en les calibrant sur les valeurs réelles de plusieurs paramètres génétiques, elles restent une simplification de la réalité, avec un déterminisme génétique purement additif, des QTL biallèliques, aucune erreur de génotypage, aucune donnée moléculaire manquante et, pour la stratégie RRGS_HYB, des phases connues sans erreurs chez les hybrides. L'intérêt de la SG par rapport à la SRR classique est à rechercher dans le ratio $r_{\hat{A},A} \times \sigma_a \times i / L$ respectif de ces deux stratégies d'amélioration génétique. Les simulations ont montré que le paramètre clé de la SG est la précision des GEBV des individus non testés en croisement car c'est lui qui détermine s'il est possible de se passer des tests en croisement et qui conditionne finalement L et i. A la suite de la thèse, le principal travail à accomplir avant de pouvoir appliquer la SG chez le palmier à huile est donc d'obtenir une estimation empirique du rythme de réduction de la précision de la SG au fil des générations sans recalibration du modèle, pour les deux groupes parentaux et toutes les composantes du rendement. Une première réponse sera apportée prochainement par l'analyse du nouveau jeu de données (Aek Loba et Aek Kwasan 2).

Le Chapitre IV a confirmé la faible diversité génétique existante au sein des deux groupes parentaux. Il apparait nécessaire de les élargir en faisant par exemple un usage plus important des Angola (groupe A) et en recourant à d'autres populations africaines (Cameroun, Ghana, etc.). L'investissement que cela peut représenter dans un premier temps pour mettre à niveau ces populations (qui n'ont pas été autant améliorées par le passé que les Deli, les la Mé ou les Yangambi) devrait se trouver largement justifié par les progrès génétiques qu'ils devraient permettre sur le long terme, en apportant vraisemblablement de la variabilité à certains gènes d'intérêt fixés dans les populations actuelles ou en apportant de nouveaux allèles à des gènes pour lesquels il existe déjà de la variabilité. Dans ce domaine, il semble pour l'instant impossible d'avoir recours directement à la SG. En effet, une autre conclusion de l'étude empirique de la SG (Chapitre V) est l'impossibilité de l'appliquer à des familles faiblement apparentées au matériel courant du programme d'amélioration (c-à-d à la population d'apprentissage) puisque la précision obtenue est alors très faible, voire négative. Dans la pratique, on pourra déterminer si les nouvelles familles (obtenues par exemple grâce à un programme d'échanges entre sociétés semencières de palmier à huile) sont suffisamment apparentées à la population d'apprentissage en calculant l'a_{max} qui les sépare. Lorsque celui-ci est important, l'intégration de ces nouvelles familles au schéma de SG devra passer par une phase de test sur descendance avec le matériel courant, afin de pouvoir éventuellement les inclure ensuite à la population d'apprentissage.

Les simulations ont montré qu'il était important d'adopter une politique de gestion de la consanguinité dans les populations d'amélioration. Ceci est d'autant plus important que la diversité génétique est déjà faible. Dans les deux premiers cycles de SRR, les sélectionneurs ont parfois essayé de minimiser la consanguinité mais de manière simple, en privilégiant par

exemple des croisements entre familles différentes plutôt qu'entre plein-frères. Cependant, cette approche n'est pas optimale, et avec l'avancement des générations et le plus grand nombre de croisements intra-groupes, elle deviendra délicate à appliquer. Il faudra donc adopter des méthodes du type (*genomic*) *optimal contribution selection* et / ou *mate selection* (discutées au Chapitre VI) pour gérer la consanguinité et minimiser son accroissement au fil des générations. Le code écrit pour les simulations réalisées au Chapitre VI pourrait facilement être adapté pour étudier cet aspect.

Enfin, comme il a été discuté à la fin du Chapitre V, des raisons pratiques ont poussé à réaliser l'étude sur les données réelles en deux étapes, avec une première étape d'estimation des AGC des parents puis une seconde étape d'utilisation en SG pour calibrer les modèles et pour estimer la précision des GEBV des candidats à la sélection. En conditions réelles, la prédiction des GEBV des candidats serait faite en une seule étape. Compte tenu de l'absence d'effet de la méthode statistique de SG mise en évidence dans le Chapitre V, on opterait certainement pour le GBLUP. En effet, il est capable de prendre en compte toute la complexité des dispositifs de tests en croisement, s'implémente avec des logiciels standards et tourne rapidement. Par ailleurs, sa mise en pratique serait directe puisqu'il s'agirait simplement d'appliquer le modèle actuel d'analyse des essais génétiques (III. A. 2. a) en remplaçant les matrices d'apparentement généalogique par des matrices moléculaires. C'est d'ailleurs cette option qui a été retenue dans le Chapitre VI pour la SG, dans une version simplifiée puisque les tests en croisements simulés n'incluaient pas tous les effets présents dans les dispositifs réels (essais, blocs, etc.) et parce que seuls des effets additifs (AGC) étaient considérés dans le modèle.

VII. B. La sélection génomique récurrente réciproque : un nouveau schéma d'amélioration pour le palmier à huile

VII. B. 1. Organisation pratique et avantages

Les résultats obtenus laissent imaginer un nouveau schéma d'amélioration pour le rendement du palmier à huile, plus performant, que l'on pourrait qualifier de sélection génomique récurrente réciproque (SGRR). Il est présenté dans la Figure 36. Les chiffres sont donnés à titre d'exemple pour indiquer un ordre de grandeur. Ils correspondent à un programme d'amélioration réaliste de SRR classique comportant 1 500 individus par groupe parental et par génération (Figure 36, à gauche). Ceux-ci constituent les champs semenciers, utilisés pour produire des semences hybrides intergroupes (étape A) et sont soumis, au moins pour une partie d'entre eux, à une présélection sur valeurs propres pour les caractères les plus héritables afin d'identifier par groupe parental 150 individus à tester en croisement (B, C). A l'issue des tests, 20 individus sont sélectionnés par groupe parental (D) et sont utilisés pour produire la génération suivante (E). Par ailleurs, une sélection phénotypique peut s'effectuer au sein des meilleurs croisements intergroupes afin d'identifier des ortets (F), utilisés pour une sortie variétale clonale. Ce type de matériel végétal n'a pas été considéré dans cette thèse car en termes de quantité produite il est secondaire par rapport aux semences. Il est toutefois

intéressant car il permet un gain génétique supplémentaire en valorisant la ségrégation génétique et les effets non additifs qui existent au sein des croisements hybrides. Par rapport à ce schéma classique, la SGRR (Figure 36, à droite) :

- raccourcit l'intervalle de génération en ne faisant des tests en croisements qu'à une génération sur deux, au lieu de chaque génération. Le temps nécessaire pour réaliser deux cycles d'amélioration est surtout influencé par les conditions environnementales. Par exemple le programme avance moins vite au Bénin, où le déficit hydrique rend les palmiers moins productifs, qu'en Asie. En moyenne, on peut considérer que 24 ans seraient nécessaires pour deux cycles avec la SGRR, contre 40 avec la SRR, soit une réduction de 40%.
- augmente l'intensité de sélection et la valeur additive moyenne des champs semenciers. En effet dans la SRR, la sélection s'opère uniquement parmi les individus testés en croisement alors qu'avec la SGRR, la sélection s'opère parmi tous les individus génotypés, testés en croisement ou non (étape D). Dans cet exemple, la sélection dans la génération servant de point de départ à l'application de la SGRR peut se faire parmi un ensemble pouvant compter jusqu'à 1 500 individus, soit 10 fois plus que la SRR. Dans la pratique, le nombre d'individus à génotyper est à raisonner en fonction des coûts liés au génotypage et, au sein de chaque groupe, de la variance des caractères d'intérêt (une forte variance justifiant que l'on génotype un plus grand nombre d'individus). Dans les générations suivantes, la SG offre un intérêt supplémentaire, celui d'accroître la valeur génétique additive moyenne du champ semencier, un aspect qui n'a pas été considéré dans cette thèse. Dans la SRR chaque nouvelle génération est constituée d'un échantillon aléatoire des descendants des individus sélectionnés de la génération précédente. En moyenne, leur valeur génétique additive est égale à celle des individus sélectionnés. Cependant, la mise en place d'un champ semencier de 1 500 individus passe par la production d'un nombre beaucoup plus grand de plantules (3 000 dans notre exemple), actuellement inutiles. Avec la SGRR, on peut génotyper ces plantules et prédire leur GEBV (F) afin de ne planter que les 1 500 meilleures (G). La valeur génétique additive moyenne des champs semenciers serait donc supérieure à celle des parents sélectionnés. Les 20 individus qui serviront à produire la génération suivante, c-à-d les 20 meilleurs parmi les 1 500 plantés (H), seront identifiés à cette même étape.
- devrait permettre d'identifier plus précisément les ortets hybrides : actuellement, ils sont sélectionnés sur la base de l'estimation de leur valeur génétique totale, c-à-d leur valeur propre corrigée des effets pris en compte par le dispositif expérimental. Cependant, cette valeur est peu précise, en particulier pour les composantes de la qualité du régime pour lesquelles peu de mesures sont faites par individu hybride (en moyenne 2.6 à Aek Loba, avec une gamme de 1 à 8) et il est difficile de faire la part des choses entre les effets génétiques propres à un individu et son micro-environnement. La SG devrait permettre d'estimer plus précisément les valeurs génétiques totales individuelles, même s'il faudra pour cela utiliser un modèle un peu différent puisqu'il devra aussi prendre en compte les effets non additifs. Son efficacité dépendra de sa capacité à tenir compte des termes de ségrégation mendélienne additive et de dominance au sein des familles de plein-frères hybrides.

Chez le palmier à huile, la SG présente un autre intérêt par rapport à la SRR, qui n'a pas été évalué dans cette thèse. Dans la SRR, les individus A et B qui sont testés en croisement sont d'abord présélectionnés sur leur valeur propre pour les composantes du rendement les plus héritables, soit essentiellement %PF et %HP. Cependant, pour des raisons pratiques, tous les individus qui seront testés en croisement ne sont pas soumis à cette présélection. Ainsi, dans le groupe B les pisifera n'ont pas de valeur propre pour les composantes du rendement et ne sont donc pas présélectionnés alors que les tenera le sont souvent. Par ailleurs, la présélection peut aussi se faire uniquement dans certaines familles, soit pour des raisons de temps soit parce que les autres familles sont dans des conditions ne permettant pas d'obtenir des valeurs propres fiables. Or, l'analyse BLUP des tests sur descendance suppose que les termes de ségrégation mendélienne soient aléatoires, alors que les individus qui seront testés en croisement après avoir été présélectionnés auront des termes de ségrégation mendélienne plus élevés que les autres. A l'issue des tests en croisements, cela abouti à des valeurs additives sous-estimées et moins précises chez les individus présélectionnés par rapport aux individus choisis au hasard (Patry et Ducrocq, 2011). La SG pourrait résoudre ce problème, à condition que les parents des tests en croisements soient choisis au hasard. Dans le cadre de la SG ceci ne serait pas gênant puisque le but des essais génétiques serait de calibrer le modèle de SG pour sélectionner parmi les parents des tests et leurs collatéraux (non testés) de la même génération, et non plus de sélectionner directement parmi les parents des tests. Dans ces conditions, choisir au hasard les individus à tester en croisement ne serait pas dommageable en termes de gain génétique. Une étude reste cependant nécessaire pour savoir s'il vaut mieux faire ce choix au hasard ou par la méthode d'optimisation CDmean. Ceci peut être fait en adaptant le code écrit pour les simulations réalisées au Chapitre VI.

Comme cela a été souligné au Chapitre V et au Chapitre VI, la meilleure façon d'atteindre une population d'apprentissage significativement plus grande que ce qui a été considéré dans cette thèse sera d'utiliser conjointement les données des parents de plusieurs cycles d'amélioration génétique. Cependant, on ne peut pas envisager de génotyper les parents d'une génération et de ne conserver que ces données en espérant, des années plus tard, les agglomérer aux données moléculaires de la nouvelle génération. En effet il est vraisemblable qu'entre temps, compte tenu de l'évolution rapide des technologies de marquage moléculaire, les données soient de type différent (SSR, SNP, séquence, etc.). Il faudra donc prélever et stocker des quantités importantes de feuilles et / ou d'ADN sur le long-terme, afin de pouvoir, à chaque génération de calibration du modèle de GS, génotyper l'ensemble des individus, toutes générations confondues.

La possibilité d'utiliser la SG pour obtenir des AGC plus précises pour les individus testés en croisement a été étudiée avec le GBLUP et les données réelles pour NR et PM dans le cadre d'un stage de M2 encadré pendant cette thèse (Marchal, 2014). Etant donné que le modèle mixte généalogique et le modèle mixte génomique (GBLUP) ne renvoient pas aux mêmes populations de base, il n'est pas pertinent de comparer leurs précisions calculées avec la formule des PEV (éq. [14]). La comparaison a donc porté sur la vraisemblance des modèles, ce qui revenait dans notre cas à comparer leurs critères de vraisemblance pénalisés

(critère d'information d'Akaïké (AIC) et critère d'information Bayésien (BIC)). Elle a montré que le GBLUP avait une plus grande vraisemblance que le BLUP généalogique. Cependant, les corrélations entre les AGC estimées par les deux modèles étaient très fortes (>0.95), indiquant que l'amélioration apportée par le GBLUP par rapport au BLUP classique était marginale. D'ailleurs, on voit dans les simulations que pour les individus testés en croisement, la précision du BLUP classique est similaire à celle du GBLUP, au moins pour la stratégie RRGS_PAR. Cet aspect pourrait toutefois être étudié un peu plus en détail. Dans la SRR, le BLUP peut uniquement tenir compte du terme de ségrégation mendélienne d'un individu par le phénotype de ses descendants de tests en croisement. Dans la SG, le modèle dispose d'une information supplémentaire pour estimer le terme de ségrégation mendélienne, les apparentements réalisés (et non pas attendus). Il serait intéressant d'étudier si cette information supplémentaire pourrait permettre de diminuer le nombre de répétitions par essai sans perte de précision pour les individus testés en croisement (par rapport aux précisions actuelles). Cette diminution du nombre de répétitions rendrait possible, pour un effort d'évaluation constant sur le terrain, de tester en croisement plus de parents et donc d'avoir une population d'apprentissage plus grande. Ce point peut être étudié avec les données réelles disponibles.

VII. B. 2. Gain génétique par unité de coût

Cette thèse n'a pas pris en compte le coût des différentes stratégies d'amélioration, qui est pourtant un point capital. Dans l'idéal, les comparaisons entre stratégies d'amélioration devraient se faire à coût constant. La RRGS génère des coûts supplémentaires liés au génotypage de la population d'apprentissage et des candidats à la sélection, c-à-d le prélèvement des échantillons de feuilles, l'extraction de l'ADN, le génotypage et le traitement des données moléculaires. La stratégie RRGS_HYB requiert beaucoup plus de génotypage que la stratégie RRGS_PAR à cause du génotypage des hybrides. Cependant, ceci ne générera pas forcément des coûts supplémentaires. En effet, des échantillons foliaires sont déjà collectés sur un échantillon d'individus au sein de chaque croisement planté en essai, afin d'en vérifier la légitimité avec un jeu de SSR (Durand-Gasselin et al., 2009). Par conséquent, en utilisant ces mêmes individus pour la calibration de la sélection génomique, le seul coût supplémentaire de la RRGS_HYB par rapport à la RRGS_PAR serait celui du génotypage SNP proprement dit des hybrides. Dans l'idéal, le développement d'une nouvelle méthode de contrôle de la légitimité basée sur des SNP permettrait d'avoir des coûts similaires pour la RRGS_HYB et la RRGS_PAR.

Cependant, les tests en croisement étant couteux (en particulier à cause de la main d'œuvre nécessaire à l'acquisition des données phénotypiques), un schéma de RRGS permettant de réaliser des tests en croisement seulement une génération sur deux générera aussi des économies, qui pourraient compenser les coûts liés au génotypage. En 2014, on pouvait obtenir le génotypage par GBS d'un individu, depuis l'extraction de l'ADN à partir d'échantillons de feuilles jusqu'à la réception des génotypes, pour 60 à 80 euros. Même en incluant les autres frais (collecte et préparation des échantillons foliaires, analyse statistique des données pour produire les GEBV), ce coût restera largement inférieur au coût de

l'évaluation en croisement d'un individu au champ. Par conséquent, il devrait être possible de passer de la SRR à la RRGS en gardant des coûts constants, uniquement en prenant comme variable d'ajustement le nombre de candidats génotypés par génération.

Clairement, une étude est nécessaire pour estimer le coût de l'évaluation en croisement d'un individu et l'ensemble des coûts liés au génotypage.

VII. B. 3. Choix de la méthode de génotypage

Le nombre de marqueurs utilisés ici sur le jeu de données réelles était réduit. Toutes les études empiriques de SG à l'exception de celle-ci ont utilisé des SNP. En effet, les génotypages sont plus faciles avec les SNP qu'avec les SSR, ce qui permet d'avoir une densité de marquage plus grande. Pour cette thèse, le choix des marqueurs moléculaires s'est porté sur les SSR car au moment de démarrer en 2009, seuls des SSR étaient disponibles pour le palmier à huile et il n'y avait pas de perspective de disposer rapidement d'un nombre satisfaisant de SNP.

Aujourd'hui, grâce aux progrès technologiques réalisés dans ce domaine et en particulier avec le développement du séquençage de nouvelle génération (NGS, pour *next generation sequencing*) (Schadt et al., 2010; van Dijk et al., 2014), des alternatives rendent accessibles à toutes les espèces un génotypage haut débit, en particulier les puces à ADN et le génotypage par séquençage (GBS, pour *genotyping by sequencing*), décrites dans l'Encadré 1.

Chez le palmier à huile, une puce à ADN a été mise au point récemment (Ting et al., 2014). La stratégie adoptée reposait sur un séquençage ciblé, limité aux zones hypométhylées (riches en gènes) du génome de huit individus. Ceci a abouti à la création d'une puce de 4 451 SNP (OPSNP3). Les SNP retenus ont une position unique et offrent une couverture régulière, sur la base de leur position dans la séquence publique du palmier à huile (Singh et al., 2013). Les puces à ADN permettent d'avoir un taux de données manquantes plus faible que le GBS, mais leur coût par échantillon est plus élevé. Poland et al. (2012) ont comparé expérimentalement la précision de la SG avec des données moléculaires obtenues par GBS et par puce à ADN chez le blé. Ils ont conclu que les deux méthodes de génotypage donnaient les mêmes précisions de SG, et que le GBS, moins couteux, était préférable. Dans plusieurs études empiriques chez le soja (Jarquín, Kocak, et al., 2014), le maïs (Crossa et al., 2013) et le manioc (Ly et al., 2013), le GBS est aussi apparu comme une méthode de génotypage adaptée pour la SG. En conséquence, en 2014, un ensemble de 214 Deli et de 197 individus du groupe B ont été génotypés par DArTseq. Ceci correspond aux parents testés en croisements à Aek Loba (c-à-d les individus utilisés dans le Chapitre V) et ceux en cours de test à Aek Kwasan 2 (Indonésie). Les individus testés à Aek Kwasan 2 se répartissent entre des individus déjà testés à Aek Loba, des individus des mêmes familles et de la même génération mais jamais testés en croisement, des individus de nouvelles familles et des descendants d'individus sélectionnés à Aek Loba. Plusieurs milliers de SNP ont été obtenus, offrant de nombreuses possibilités en matière d'évaluation de stratégies de sélection génomique et de caractérisation des populations d'amélioration.

VII. B. 4. Sélection multicaractères

Il existe des corrélations génétiques entre certaines composantes du rendement : entre NR et PM, entre %PF et %KF (forte corrélation négative dans les deux cas) et probablement ailleurs, par exemple entre NF et PF. L'existence d'une corrélation peut être exploitée dans les modèles d'analyse génétique, génomiques ou généalogiques, dans le but d'augmenter la précision de la sélection (voir Jia et Jannink (2012), Hayashi et Iwata (2013) et Guo et al. (2014) pour des modèles de SG multicaractères). L'estimation des AGC pour NR et PM au Chapitre IV a été faite avec un modèle mixte (généalogique) bivarié, ce qui a effectivement abouti à une augmentation, faible mais significative, de la précision des AGC (pas montré). Ce résultat a été confirmé avec un modèle génomique (GBLUP) dans une étude menée dans le cadre d'un stage de M2, sur le jeu de données du Chapitre IV (Marchal, 2014). Il serait intéressant d'étendre l'approche bivariée à toutes les composantes corrélées.

En plus du rendement en huile de palme, il existe un second caractère fondamental pour l'amélioration génétique du palmier à huile, la résistance aux maladies. Trois maladies majeures existent : la fusariose, la pourriture basale du stipe due au ganoderma et la pourriture du cœur. L'essentiel des zones de production étant soumises à une maladie majeure, il est indispensable d'associer rendement élevé et résistance à une des maladies. On s'attend aussi à ce que ces maladies se retrouvent à l'avenir de plus en plus souvent associées, sous l'effet des connexions entre pays producteurs et de la succession de générations de culture sur les mêmes zones, rendant les résistances multiples nécessaires.

L'évaluation de la résistance aux maladies est indépendante des tests en croisement destinés à sélectionner pour le rendement. Elle passe actuellement par des tests en pépinière avec inoculation artificielle du pathogène (fusariose et ganoderma) ou par des essais au champ (pourriture du cœur). Ces évaluations sont longues et complexes, si bien que les programmes d'évaluation pour le rendement et pour la résistance aux maladies progressent à des rythmes comparables. En faisant évoluer la sélection pour le rendement vers la RRGS un décalage apparaitra, ce qui rendra difficile la sélection combinée pour le rendement et la résistance aux maladies. La meilleure solution serait de développer des modèles de sélection génomique efficaces pour la résistance aux maladies afin d'avoir une approche génomique globale. Pour la fusariose et le ganoderma, il est possible d'initier une étude expérimentale visant à estimer la précision de la sélection génomique pour la résistance à ces deux maladies, en s'inspirant de la méthode qui a été suivie au Chapitre V. Pour la pourriture du cœur, cela n'est pas encore possible. En effet, cette maladie est toujours mal connue (l'agent pathogène n'a pas été identifié avec certitude) et les premières expérimentations visant à évaluer le niveau de résistance de géniteurs sont très récentes. Les résultats obtenus ici sur les composantes du rendement ne seront pas forcément généralisables aux résistances aux maladies, qui pourraient avoir des architectures génétiques combinant quelques QTL majeurs et des QTL mineurs. Ceci pourrait avoir un effet sur la précision de la sélection génomique et peut être faire apparaitre des différences en fonction de la méthode statistique. En attendant de pouvoir conduire une sélection génomique pour le rendement et pour la résistance aux maladies, il est possible de démarrer en appliquant la sélection génomique pour le rendement dans des familles présentant un niveau de résistance élevé à au moins une maladie.

Un seul caractère végétatif fait l'objet d'une sélection en routine, la vitesse de croissance en hauteur. Chez le palmier à huile, le but est de la ralentir afin de faciliter la récolte. La sélection pour une répartition plus régulière des récoltes sur l'année est aussi un objectif intéressant, mais qui a été peu étudié jusqu'à présent faute d'un indicateur pour mesurer cette répartition. Récemment, l'indice de Gini a été proposé pour combler ce manque (Cros et al., 2013). Une étude menée pendant cette thèse dans le cadre d'un stage de M2 a montré, sur le jeu de données du Chapitre IV, que cet indice était pertinent pour sélectionner des parents dans le but de rendre plus régulière la production des hybrides (Souchard, 2013). Les observations pour ces deux caractères étant faites dans les mêmes essais que l'évaluation pour le rendement, on pourrait donc leur appliquer l'approche RRGS telle qu'elle a été décrite (Figure 36).

Un inconvénient possible de la sélection précoce d'individus permise par la SG est la diffusion dans la population d'amélioration de défauts génétiques qui n'auraient pas été observés avant l'étape de sélection. Ceci peut se produire pour des caractères qui ne sont pas observés en routine et qui ne peuvent donc pas être inclus dans le modèle de SG. Chez le palmier à huile, cela pourrait concerner par exemple des défauts sur les régimes, tels que la présence d'épines développées au niveau des épillets, rendant la récolte difficile. Il faudra donc systématiser l'enregistrement de ces défauts dans les populations parentales, afin de pouvoir éventuellement éliminer les descendants d'individus présentant ce type de problèmes. Il serait aussi prudent d'augmenter le nombre d'individus sélectionnés par génération par rapport à la SRR, en visant un compromis entre forte intensité de sélection et impact réduit dans le cas où un individu aurait transmis un défaut non observé et non prédit par la SG. Ceci contribuerait aussi à réduire le taux d'accroissement de la consanguinité.

VII. C. Vers des modèles de SG plus complets

VII. C. 1. Prise en compte de la structure des populations parentales

Dans le jeu de données expérimentales, les croisements étaient des hybrides intergroupes (A × B) et le modèle utilisé pour estimer les AGC, c-à-d le modèle de Stuber et Cockerham (équation [18]), considérait que chaque groupe possédait une variance additive spécifique. Cependant, les groupes sont constitués de populations initialement indépendantes. Il pourrait donc être plus juste d'associer une variance additive spécifique à chaque population plutôt qu'à chaque groupe. Par ailleurs, à partir du cycle suivant d'amélioration (c-à-d dans les tests en croisements d'Aek Kwasan 2, dans le nouveau jeu de données) des croisements de type (A × B) × B ont été évalués, pour lesquels le modèle de Stuber et Cockerham n'est pas approprié. Pour faire face à ce problème il existe un modèle multi-population, le *multibreed model* utilisé chez les animaux (Elzo, 1990; Lo et al., 1993). Cependant, dans sa version originale les composantes de la variance sont difficiles à estimer. García-Cortés et Toro (2006) ont développé un modèle équivalent simple à analyser. Celui-ci décompose la valeur additive des individus en des effets indépendants (a_p) associés chacun à une des populations initiales. Pour chaque effet, la covariance entre les individus est le produit d'une matrice

partielle d'apparentement additif ($\boldsymbol{A_p}$) et de la variance additive de la population ($\sigma^2_{a(p)}$). On peut aussi inclure des effets ($a_{pp'}$) tenant compte de la ségrégation. Il s'écrit :

$$y = \boldsymbol{X}b + \sum_{p=1}^{n_p} \boldsymbol{Z_p} a_p + \sum_{p=1}^{n_p-1} \sum_{p'=p+1}^{n_p} \boldsymbol{Z_{pp'}} a_{pp'} + e$$

avec y le vecteur des observations, a_p le vecteur des effets additifs dus aux allèles de la population p [~N(0, $\boldsymbol{A_p}\sigma^2_{a(p)}$)], $a_{pp'}$ le vecteur des effets additifs correspondant à la déviation de ségrégation entre les populations p et p' [~N(0, $\boldsymbol{A_{pp'}}\sigma^2_{a(pp')}$)], b le vecteur des effets fixes, $\boldsymbol{X}$, $\boldsymbol{Z_p}$ et $\boldsymbol{Z_{pp'}}$ des matrices d'incidence, e la résiduelle et n_p le nombre de populations initiales. Sous cette forme, le modèle peut être traité de manière standard, par exemple avec ASReml. $\boldsymbol{A_p}$ et $\boldsymbol{A_{pp'}}$ sont des matrices symétriques carrées de dimension égale au nombre d'individus présents dans le pédigrée des individus observés. Elles se construisent à partir de la proportion de gènes que chaque individu a hérité des populations p et p' et des coefficients d'apparentement généalogique. Strandén et Mäntysaari (2013) ont développé une approximation de ce modèle permettant de l'utiliser avec des données moléculaires (Makgahlela et al., 2013). Dans le cas du palmier à huile, ce modèle serait appliqué aux données phénotypiques des individus hybrides et la valeur additive en croisement des parents s'obtiendrait grâce à leurs $\hat{a}_p$ et $\hat{a}_{pp'}$, avec la formule :

$$a = \sum_{p=1}^{n_p} a_p + \sum_{p=1}^{n_p-1} \sum_{p'=p+1}^{n_p} a_{pp'}.$$

VII. C. 2. Prise en compte d'effets non additifs

Le modèle mixte classique estimant les AGC des palmiers à huile testés en croisement inclut un effet d'ASC, qui permet de voir que même si les composantes du rendement sont globalement additives, il existe une part de dominance, plus ou moins forte selon le caractère. De la même manière, une part d'épistasie pourrait probablement être mise en évidence. Il serait intéressant d'inclure des effets non additifs au modèle de SG destiné à analyser les tests en croisement, à la fois pour sélectionner les meilleurs parents et, comme on l'a envisagé précédemment, les ortets. Plusieurs études ont porté sur l'incorporation de tels effets dans les modèles de SG.

Avec le RR-BLUP, l'ajout d'un terme de dominance est simple. En prenant comme exemple une population d'apprentissage composée d'individus hybrides interpopulations A × B, on peut utiliser un modèle avec des effets aux marqueurs additifs (spécifiques à la population parentale) et un terme de dominance, comme chez le maïs dans Technow et al. (2012) et Massman et al. (2013) :

$$y = 1\mu + \boldsymbol{Z_A}\, m_A + \boldsymbol{Z_B}\, m_B + \boldsymbol{Z_d}\, d_{AB} + e$$

où m_A et m_B sont les vecteurs des effets additifs associés à chacun des SNP pour les populations parentales A et B (n_{SNP} × 1), $\boldsymbol{Z_A}$ et $\boldsymbol{Z_B}$ sont les matrices d'incidence des effets additifs, d_{AB} est le vecteur de l'effet de dominance associé à chaque SNP (n_{SNP} × 1) et $\boldsymbol{Z_d}$ est une matrice d'incidence dont les éléments Z_{dij} valent 1 si l'individu i est hétérozygote au SNP j et 0 s'il est homozygote ($n_{individu}$ × n_{SNP}).

On peut aussi calculer une matrice de dominance $\boldsymbol{D}$ à partir de données moléculaires (Su et al., 2012; Vitezica et al., 2013; Da et al., 2014) et l'utiliser dans un modèle mixte génomique (GBLUP) en complément de la matrice d'apparentement additif génomique de

manière à estimer conjointement la valeur additive et la valeur de dominance des individus génotypés. Chez des bovins laitiers, Sun et al. (2014) ont montré que la précision du GBLUP pour les caractères liés à la production augmentait lorsque le modèle comportait à la fois des effets additifs et de dominance. Su et al. (2012) a aussi estimé des effets d'épistasie, en utilisant le fait que la matrice d'apparentement épistatique $\boldsymbol{G_{aa}}$ est approximativement égale au produit de Hadamard (produit terme à terme) de la matrice d'apparentement additive (généalogique ou moléculaire) avec elle même. Chez le porc, ils ont montré que le meilleur modèle en termes de précision et de biais de sélection était le modèle incluant à la fois les effets additifs, de dominance et d'épistasie.

On peut aussi inclure un effet de dominance dans les méthodes statistiques bayésiennes de SG. On trouve des exemples avec des applications sur des données réelles et simulées dans Denis et Bouvet (2013) chez l'eucalyptus, Zhao et al. (2013) chez le blé et Technow et al. (2012).

Plusieurs études se sont penchées sur le cas des croisements hybrides interpopulations et leurs résultats sont donc particulièrement intéressants pour le palmier à huile. Elles ont montré que l'utilisation d'un terme de dominance dans le modèle de SG pouvait améliorer la précision des GEBV. Zeng et al. (2013) ont simulé un programme hybride entre deux populations dans lesquelles ils ont appliqué de la SG avec des modèles bayésiens avec ou sans dominance, calibrés sur les hybrides. Ils montré que plus la part de dominance dans le déterminisme génétique du caractère était importante, plus la réponse à la sélection du modèle avec dominance dépassait celle du modèle additif. En l'absence de dominance, les deux modèles avaient des performances similaires. Technow et al. (2012) ont comparé par simulation sur des hybrides interpopulations de maïs la précision du RR-BLUP et de BayesB pour des modèles avec ou sans dominance. Ils ont trouvé que la présence d'un terme de dominance améliorait considérablement la précision lorsque les populations parentales avaient des fréquences allèliques proches (part élevée de la variance de dominance dans la variance génétique totale entre hybrides). Ils ont aussi montré que la précision des modèles avec dominance était significativement plus grande avec BayesB qu'avec le GBLUP. Cependant, Zhao et al. (2013) en appliquant le RR-BLUP et des méthodes bayésiennes à des données réelles et simulées d'hybrides de blé ont montré qu'ajouter un terme de dominance n'améliorait pas la précision, et pouvait parfois la réduire.

VII. C. 3. Prise en compte d'informations a priori sur l'effet des marqueurs

Zhang et al. (2010) ont proposé un modèle GBLUP amélioré, le TABLUP. Celui-ci diffère du GBLUP car il exploite les résultats de méthodes de SG qui estiment des effets aux marqueurs pour construire une matrice d'apparentement additif réalisé spécifique au caractère considéré (matrice $\boldsymbol{TA}$). Pour des marqueurs biallèliques, on définit une matrice diagonale $\boldsymbol{W}$ contenant pour chaque marqueur m un poids w_m correspondant à la proportion de la variance additive totale expliquée par le marqueur, estimée par BayesB (TABLUP_B) ou RR-BLUP (TABLUP_RR) : $w_m = \sigma^2_{a_m} \,/\, \sum_{i=1}^{N_m} \sigma^2_{a_i}$, avec $\sigma^2_{a_m}$ la variance additive du marqueur m et N_m le nombre total de marqueurs. Avec BayesB, les variances additives des marqueurs font partie

des résultats du modèle. Avec le RR-BLUP, on les obtient par la formule $\sigma^2_{a_m} = 2p_m(1 - p_m)\hat{a}^2_m$, avec $\hat{a}^2_m$ l'estimation de l'effet du marqueur m et p_m la fréquence de l'un de ses allèles (cf II. C). La matrice $\boldsymbol{TA}$ vaut (Zhang et al., 2014) :

$$\boldsymbol{TA} = \frac{\boldsymbol{X}\,\boldsymbol{W}\,{}^{\boldsymbol{t}}(\boldsymbol{X})}{2\sum_{i=1}^{N_m} p_i(1-p_i)}$$

avec $\boldsymbol{X} = \boldsymbol{Z} - \boldsymbol{P}$, $\boldsymbol{Z}$ codée en 0, 1 ou 2 selon le génotype et $\boldsymbol{P}$ une matrice avec les individus en ligne, les marqueurs en colonnes et chaque colonne m remplie de $2p_m$, c'est-à-dire du double de la fréquence de l'allèle le moins fréquent du marqueur m. Dans le TABLUP, tous les marqueurs n'ont donc pas le même poids dans le calcul des apparentements. Dans des simulations (Zhang et al., 2010; Wang et al., 2012), la précision du TABLUP était supérieure à celle du BLUP traditionnel (utilisant la matrice $\boldsymbol{A}$), du GBLUP et du RR-BLUP ; et le TABLUP_B était la version la plus performante du TABLUP. Le TABLUP_B avait par contre une précision égale ou légèrement inférieure à celle des méthodes BayesA, BayesB et BayesCπ.

Le TABLUP est donc un GBLUP qui utilise pour le caractère étudié une information *a priori*, l'estimation par d'autres méthodes de SG de la variance additive expliquée par chaque marqueur. Cependant, d'autres types d'informations sur l'architecture génétique d'un caractère peuvent être disponibles, et notamment les résultats d'études de génétique d'association (GWAS). Afin de rendre l'approche TABLUP plus flexible et plus générale, Zhang et al. (2014) ont développé le BLUP|GA, pour *BLUP conditional on the genetic architecture*. Celui-ci utilise les informations disponibles pour pondérer l'effet des marqueurs dans le calcul de la matrice d'apparentement additif réalisé ($\boldsymbol{T}$). Sur la base de ces informations, les N_m marqueurs sont divisés en deux sous ensembles, M1 et M2, comportant N_{m1} et N_{m2} marqueurs, respectivement. Les marqueurs M1 reçoivent des poids importants et les marqueurs M2 des poids faibles :

$$\boldsymbol{T} = \omega\boldsymbol{S} + (1-\omega)\boldsymbol{G}$$

avec ω un poids global destiné à contrôler l'importance accordée aux marqueurs M1, $\boldsymbol{G}$ une matrice concernant tous les marqueurs ($N_m \times N_m$) et calculée selon VanRaden (2007) et Habier et al. (2007) (voir page 39) et $\boldsymbol{S}$ une matrice se rapportant aux marqueurs M1 ($N_{m1} \times N_{m1}$) telle que :

$$\boldsymbol{S} = \frac{\boldsymbol{X_1}\,\boldsymbol{W_1}\,{}^{\boldsymbol{t}}(\boldsymbol{X_1})}{2\sum_{i=1}^{N_{m1}} p_i(1-p_i)}$$

avec $\boldsymbol{X_1}$ la partie de $\boldsymbol{X}$ concernant uniquement les marqueurs M1 et $\boldsymbol{W_1}$ une matrice diagonale comportant les poids associés aux marqueur. Ceux-ci peuvent être les effets ou les variances estimés à chaque marqueur par un modèle de SG ou à des marqueurs situés dans des QTL mis en évidence par GWAS; ou le décompte du nombre d'études ayant mis en évidence un QTL incluant les marqueurs. Le TABLUP correspond donc au cas particulier du BLUP|GA où $\omega = 1$, $N_{m1} = N_m$ et avec la diagonale de $\boldsymbol{W_1}$ qui donne la proportion de la variance additive totale expliquée par chaque marqueur. Les performances du BLUP|GA ont été évaluées avec des données réelles de Holstein et de riz, en utilisant comme poids pour les marqueurs le nombre de publications de GWAS ayant mis en évidence un QTL les incluant. La précision du BLUP|GA a dépassé celle des modèles standards de SG (GBLUP et BayesB) pour deux caractères sur trois chez les Holstein et neuf caractères sur 11 chez le riz. Les résultats sur des

données simulées indiquent que le BLUP|GA est d'autant plus intéressant que la population d'apprentissage est petite (avec comme modalités testées 125, 500 et 2 000 individus).

Bernardo (2014) a montré par simulations que, lorsque quelques gènes majeurs sont impliqués dans le contrôle d'un caractère et qu'ils expliquent chacun au moins 10% de la variance génétique, leur modélisation par des effets fixes augmentait la précision. Rutkoski et al. (2014) ont confirmé empiriquement ce résultat sur la résistance quantitative à la rouille du blé. Une approche GWAS a identifié deux marqueurs expliquant 27% et 17% de la variation génotypique et étroitement liés à un gène connu pour avoir un effet modéré sur la résistance à la rouille. Ces marqueurs ont été inclus en effets fixes dans un modèle GBLUP :

$$y = \mu + \boldsymbol{X_1}\beta_1 + \boldsymbol{X_2}\beta_2 + \boldsymbol{Z_u}u + e$$

avec y le vecteur des données, μ la moyenne phénotypique, u le vecteur des GEBV associé à une matrice classique d'apparentement moléculaire $\boldsymbol{G}$, β_1 et β_2 les effets fixes des deux marqueurs fonctionnels et $\boldsymbol{X_1}$, $\boldsymbol{X_2}$ et $\boldsymbol{Z_u}$ des matrices d'incidence ($\boldsymbol{X_1}$ et $\boldsymbol{X_2}$ étant des matrices colonnes renseignant sur le génotype des individus aux deux marqueurs). Ce modèle a été comparé au GBLUP classique, au LASSO Bayésien et à BayesCπ et est apparu comme le modèle donnant la précision la plus élevée.

Le MultiBLUP (Speed et Balding, 2014) est une extension du RR-BLUP permettant de prendre en compte des classes de marqueurs qui diffèrent par l'amplitude supposée de leurs effets. Elles peuvent être définies de nombreuses manières : marqueurs situés dans les exons, les introns ou flanquants, fonctionnels ou anonymes, dans des QTL ou non, etc. Pour un ensemble de N_m marqueurs répartis entre M classes, le modèle MultiBLUP s'écrit :

$$y = \mu + \sum_{i=1}^{M} \boldsymbol{Z_i}\, m_i + e$$

avec m_i le vecteur de l'effet additif des marqueurs et $\boldsymbol{Z_i}$ la part de la matrice $\boldsymbol{Z}$ relative aux marqueurs i. On a alors $m_i \sim N(0, \boldsymbol{K_i}\,\sigma^2_{mi})$, avec $\boldsymbol{K_i}$ une matrice de similarité $\boldsymbol{K_i} = \boldsymbol{Z}^t\boldsymbol{(Z)} / N_m$ et σ^2_{mi} la variance des marqueurs du sous ensemble i. Le modèle estime alors une variance des marqueurs spécifique à chaque classe. Speed et Balding (2014) ont aussi développé une version « adaptative » du MultiBLUP, qui identifie automatiquement les classes de SNP. Evalué empiriquement sur un vaste ensemble de maladies humaines, le MultiBLUP est apparu comme la meilleure méthode de SG.

Dans la même logique, le W-BLUP (Zhao et al., 2013) ou *weighted BLUP* est une autre extension du RR-BLUP destinée à distinguer des marqueurs anonymes et des marqueurs fonctionnels. Il a été développé sous une forme estimant des effets aux marqueurs additifs et de dominance et appliqué sur le blé. Trois marqueurs fonctionnels ont été utilisés, situés chacun dans un gène impliqué dans le contrôle des caractères étudiés. Le modèle W-BLUP était de la forme :

$$y = \mu + \boldsymbol{Z_A}a + \boldsymbol{F_A}a_f + \boldsymbol{Z_D}d + \boldsymbol{F_D}d_f + e$$

avec a et d les vecteurs contenant respectivement les effets additifs et de dominance associés aux marqueurs anonymes, a_f et d_f les vecteurs des effets additifs et de dominance (aléatoires) associés aux marqueurs fonctionnels et $\boldsymbol{Z_A}$, $\boldsymbol{F_A}$, $\boldsymbol{Z_D}$ et $\boldsymbol{F_D}$ des matrices d'incidence renseignant sur le génotype des marqueurs considérés. La séparation des marqueurs anonymes et fonctionnels en des termes distincts permet de leur accorder une importance différente. Sur

des données expérimentales, le W-BLUP a donné les précisions les plus fortes, comparé au RR-BLUP et à BayesCπ.

Ces modèles de SG perfectionnés semblent donc des alternatives intéressantes à évaluer pour le palmier à huile, en particulier avec le nouveau jeu de données dont le plus grand nombre d'individus et de marqueurs pourraient contribuer à faire apparaitre des différences entre modèles. Des informations a priori sur certains marqueurs pourront être obtenues grâce aux études de transcriptomique (Ho et al., 2007; Tranbarger et al., 2011; Dussert et al., 2013) et de détection de QTL (voir III. A. 3. d) qui ont été faites sur des composantes du rendement et grâce à la séquence publique (Singh et al., 2013).

VII. C. 4. Prise en compte d'autres données « -omiques »

En dehors des données génomiques, d'autres types d'informations peuvent être inclus dans les modèles de SG afin d'en améliorer la précision, en particulier les résultats d'études de transcriptomique, de protéomique et de métabolomique. Riedelsheimer et al. (2012) ont combiné des données génomiques et métabolomiques[4] pour prédire les aptitudes à la combinaison chez le maïs, avec un modèle de SG utilisant des SNP (+50K) et des métabolites (130) et ont conclu que l'utilisation des métabolites permettait d'obtenir de bonnes précisions. Cependant, l'application pratique de la SG implique de disposer des données entrant dans le modèle prédictif pour les individus de la population d'apprentissage et pour les candidats à la sélection, ce qui chez le palmier à huile devrait représenter entre quelques centaines et quelques milliers d'individus par génération. Or actuellement il est beaucoup plus facile de produire des données génomiques en routine sur un grand nombre d'individus que d'autres types de données « -omiques ». Par ailleurs les données génomiques ont l'avantage de ne nécessiter qu'un seul prélèvement, fait à n'importe quel stade de développement de l'individu et indépendamment des conditions environnementales. Les études faites dans les autres domaines « -omiques » sont dynamiques ; et il se pose donc les questions du stade de développement auquel collecter l'information, de la fréquence de cette collecte, du tissu à considérer et de l'effet de l'environnement. A court terme il ne semble donc pas possible d'utiliser ce type de données de manière pratique dans un schéma d'amélioration. Pour l'instant, leur intérêt serait plutôt à rechercher dans l'identification de zones du génome contrôlant les caractères d'intérêt afin d'intégrer ces informations au modèle de SG, avec les méthodes innovantes décrites précédemment (BLUP|GA, MultiBLUP, etc.).

[4] La métabolomique étudie l'ensemble des métabolites, c-à-d des petites molécules telles que sucres, acides aminés, acides gras, etc., présents dans une cellule, un tissu, un organe ou un organisme à un moment et dans des conditions données.

VII. C. 5. Prise en compte de données environnementales

Bien que le palmier à huile soit cultivé uniquement dans la zone tropicale, il est confronté à des conditions environnementales contrastées (climat, sol et pratiques culturales), notamment en termes de déficit hydrique, dont la gamme va de fort dans certaines zones africaines à absent, notamment en Asie du sud-est. Des interactions entre génotype et environnement ont été mises en évidence pour le rendement en huile de palme et ses composantes (Obisesan et Fatunla, 1983; Ataga, 1993; Corley et Tinker, 2003, p.195; Rafii et al., 2012). Il serait donc intéressant de les prendre en compte dans les modèles de SG afin de de prédire la valeur additive d'un candidat à la sélection dans un environnement dans lequel il n'a pas été évalué. Cette prédiction pourrait s'obtenir soit à partir de la valeur du candidat dans d'autres environnements dans lesquels il a été testé ou à partir de la valeur d'autres candidats testés dans l'environnement considéré. Pour les sélectionneurs, cela permet de définir les zones agro-environnementales optimales des croisements élites afin d'établir des stratégies pertinentes de sélection des parents et de commercialisation des croisements. Un modèle génétique combinant les résultats de tests en croisements réalisés dans différents environnements, des informations moléculaires et des covariables environnementales peut y parvenir.

Heslot et al. (2014) ont développé un modèle de SG comprenant des covariables environnementales et un modèle de culture (c-à-d qui décrit la croissance et le développement en interaction avec les conditions environnementales). Chez le blé, ce modèle a augmenté la précision des GEBV pour des environnements absents du jeu de calibration et a réduit la variabilité de la précision. Jarquín et al. (2014) ont proposé une méthode pour modéliser les interactions entre des jeux de données moléculaires et de covariables environnementales de grandes tailles. Sur le blé, en utilisant 2 395 SNP et 68 covariables environnementales, les précisions obtenues avec leur modèle étaient plus grandes qu'en négligeant les interactions.

Cependant, chez le palmier à huile la modélisation des interactions entre génotype et environnement est rendue plus délicate par la nature pérenne de l'espèce, si bien que les effets des conditions environnementales d'une année peuvent s'exprimer plusieurs années plus tard ou s'étendre sur plusieurs années, et qu'il peut exister des interactions entre les conditions environnementales de plusieurs années. Des études sont donc nécessaires pour adapter ce type de modèles à la complexité des interactions génotype × environnement existant chez le palmier à huile.

CHAPITRE VIII. CONCLUSION GÉNÉRALE

Les données expérimentales et simulées indiquent que la sélection génomique (SG) devrait révolutionner l'amélioration génétique du rendement chez le palmier à huile, en donnant la possibilité de sélectionner uniquement sur leur génotype les individus ayant les plus fortes aptitudes à la combinaison hybride. Ceci amènerait une diminution de l'intervalle moyen de génération et un accroissement de l'intensité de sélection, aboutissant à un gain génétique annuel qui pourrait dépasser de 50% celui de la méthode de sélection traditionnelle. De nouvelles études sont requises pour compléter les résultats obtenus dans cette thèse avant l'application pratique de la SG. Le principal objectif devrait désormais être d'obtenir une confirmation expérimentale des simulations, en estimant sur plusieurs générations la précision de sélection sans recalibration du modèle. Dans la perspective de la mise en œuvre de la SG, un nouveau schéma d'amélioration génétique du palmier à huile a été proposé, la sélection génomique récurrente réciproque, qui intègre l'approche génomique au schéma actuel. Les futures recherches sur la SG appliquée au palmier à huile devront aussi porter sur les autres caractères d'intérêt, et en premier lieu sur l'amélioration de la résistance aux maladies qui, compte tenu des conditions rencontrées dans les zones de culture, doit obligatoirement aller de paire avec l'amélioration du rendement. Du point de vue méthodologique, les futures recherches devraient utiliser les nouveaux modèles qui ont été développés car ceux-ci sont potentiellement plus efficaces, notamment grâce à la prise en compte des effets non additifs ou d'informations a priori sur les effets des marqueurs. Dans un deuxième temps, des approches plus ambitieuses devraient être envisagées, comme l'utilisation de modèles de SG combinant les résultats de tests en croisements réalisés dans différents environnements, des informations moléculaires et des covariables environnementales. Ceci permettrait de prédire les aptitudes à la combinaison hybride de candidats à la sélection dans des environnements dans lesquels ils n'ont pas été évalués, une possibilité dont l'intérêt devrait grandir avec le changement climatique global. Enfin, il ressort de cette étude que, dans le contexte actuel où la production agricole doit augmenter à un rythme jamais atteint pour faire face à la forte hausse attendue de la demande alimentaire, la SG a indéniablement un rôle à jouer pour l'amélioration génétique en générale et tout particulièrement pour le palmier à huile.

Bibliographie

Ataga D., 1993. Genotype-environment interaction and stability analysis for bunch yield in the oil palm (*Elaeis guineensis* Jacq.). 48(2): 59‑63.

Beavis W.D., 1994. The power and deceit of QTL experiments: lessons from comparative QTL studies. Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference, American Seed Trade Association, Washington DC, 250‑266.

Beavis W.D., 1998. QTL analyses: Power, precision, and accuracy, pp. 145‑162 in *Molecular Dissection of Complex Traits*, Paterson, A.H., Boca Raton.

Beirnaert A. et Vanderweyen R., 1941. Contribution à l'étude génétique et biométrique des variétiés d'*Elaeis guineensis* Jacq. Publ. Inst. Nat. Etude Agron. Congo Belge. Ser. Sci., 27: 1–101.

Bénard G., 1965. Caractéristiques qualitatives du régime d'*Elaeis guineensis* Jacq. : teneur en huile de la pulpe des diverses origines et des croisements interorigines. Oléagineux, 20(3): 163‑168.

Bernardo R., 1993. Estimation of coefficient of coancestry using molecular markers in maize. Theoretical and Applied Genetics, 85(8): 1055‑1062.

Bernardo R., 1996. Best Linear Unbiased Prediction of Maize Single-Cross Performance. Crop Sci., 36(1): 50‑56.

Bernardo R., 2014. Genomewide Selection when Major Genes Are Known. Crop Sci., 54(1): 68‑75.

Billotte N., Jourjon M.F., Marseillac N., Berger A., Flori A. *et al.*, 2010. QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 120(8): 1673‑1687.

Billotte N., Marseillac N., Risterucci A.-M., Adon B., Brottier P. *et al.*, 2005. Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 110(4): 754‑765.

Bouquet A. et Juga J., 2013. Integrating genomic selection into dairy cattle breeding programmes: a review. Animal, 7(05): 705‑713.

Browning S.R. et Browning B.L., 2007. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. The American Journal of Human Genetics, 81(5): 1084‑1097.

Butler D.G., Cullis B.R., Gilmour A.R. et Gogel B.J., 2009. *Mixed models for S language environments: ASReml-R reference manual (Version 3)*. Queensland Department of Primary Industries and Fisheries, 398 p.

Caballero A., 1994. Developments in the prediction of effective population size. Heredity, 73(6): 657‑679.

Calus M.P.L., Meuwissen T.H.E., de Roos A.P.W. et Veerkamp R.F., 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics, 178(1): 553‑561.

Calus M.P.L. et Veerkamp R.F., 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics, 124(6): 362‑368.

Calus M., de Roos S. et Veerkamp R., 2009. Estimating genomic breeding values from the QTL-MAS Workshop Data using a single SNP and haplotype/IBD approach. BMC Proceedings, 3(Suppl 1): S10.

De los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K. et Cotes J.M., 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics, 182(1): 375‑385.

De los Campos G., Pérez P., Vazquez A. et Crossa J., 2013. Genome-Enabled Prediction Using the BLR (Bayesian Linear Regression) R-Package, pp. 299‑320 in *Genome-Wide Association Studies and Genomic Prediction*, édité par C. Gondro, J. van der Werf, et B. Hayes. Methods in Molecular Biology, Humana Press.

Cervantes I., Goyache F., Molina A., Valera M. et Gutiérrez J.P., 2011. Estimation of effective population size from the rate of coancestry in pedigreed populations. Journal of Animal Breeding and Genetics, 128(1): 56‑63.

Cervantes I., Pastor J.M., Gutiérrez J.P., Goyache F. et Molina A., 2011. Computing effective population size from molecular data: The case of three rare Spanish ruminant populations. Livestock Science, 138(1–3): 202‑206.

Chagné D., Crowhurst R.N., Troggio M., Davey M.W., Gilmore B. *et al.*, 2012. Genome-Wide SNP Detection, Validation, and Development of an 8K SNP Array for Apple. PLoS ONE, 7(2): e31745.

Clark S., Hickey J., Daetwyler H. et van der Werf J., 2012. The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. Genetics Selection Evolution, 44(1): 4.

Cochard B., 2008. Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (*Elaeis guineensis* Jacq.) [Thèse de Doctorat]: Montpellier SupAgro, 97-[175] p.

Cochard B., Adon B., Rekima S., Billotte N., de Chenon R. *et al.*, 2009. Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. Tree Genetics & Genomes, 5(3): 493‑504.

Comstock R.E., Robinson H.F. et Harvey P.H., 1949. A breeding procedure designed to make maximum use of both general and specific combining ability. Agron. J., 41(8): 360‑367.

Corley R., 2009. How much palm oil do we need ? Environmental Science and Policy, 12: 134‑139.

Corley R.H.V. et Lee C.H., 1992. The physiological basis for genetic improvement of oil palm in Malaysia. Euphytica, 60(3): 179‑184.

Corley R. et Tinker P., 2003. Selection and breeding, pp. 133‑199 in *The oil palm*, Blackwell Science Ltd Blackwell Publishing, Oxford.

Coster A. et Bastiaansen J., 2010. *HaploSim: R package version 1.8.4*.

Cros D., Denis M., Sánchez L., Cochard B., Flori A. *et al.*, 2014. Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 1‑14.

Cros D., Flori A., Nodichao L., Omoré A. et Nouy B., 2013. Differential response to water balance and bunch load generates diversity of bunch production profiles among oil palm crosses (*Elaeis guineensis*). Tropical Plant Biology, 6(1): 26‑36.

Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M. et Fernández J., 2014. Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population. Theoretical and Applied Genetics, 127(4): 981‑994.

Crossa J., Beyene Y., Kassa S., Pérez P., Hickey J.M. *et al.*, 2013. Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. G3: Genes|Genomes|Genetics, 3(11): 1903‑1926.

Daetwyler H.D., Calus M.P.L., Pong-Wong R., de los Campos G. et Hickey J.M., 2013. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics, 193(2): 347‑365.

Daetwyler H.D., Pong-Wong R., Villanueva B. et Woolliams J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics, 185(3): 1021‑1031.

Daetwyler H.D., Villanueva B., Bijma P. et Woolliams J.A., 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics, 124(6): 369‑376.

Da Y., Wang C., Wang S. et Hu G., 2014. Mixed Model Methods for Genomic Prediction and Variance Component Estimation of Additive and Dominance Effects Using SNP Markers. PLoS ONE, 9(1): e87666.

Demol J., Baudoin J.P., Louant B.P., Maréchal R., Mergeai G. et Otoul E., 2002. *Amélioration des plantes: Application aux principales espèces cultivées en régions tropicales*. Presses Agronomiques de Gembloux, Gembloux, Belgique, 581 p.

Denis M. et Bouvet J.-M., 2013. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genetics & Genomes, 9(1): 37‑51.

Desta Z.A. et Ortiz R., 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science, (0):

Van Dijk E.L., Auger H., Jaszczyszyn Y. et Thermes C., 2014. Ten years of next-generation sequencing technology. Trends in Genetics, 30(9): 418‑426.

Durand-Gasselin T., Billotte N., Pomiès V., Mastin G., Potier F., Amblard P., Flori A. et Cochard B., 2009. ID Checking by Microsatellite Type Markers (SSR) During the oil Palm Variety Selection and Production Processes, pp. 8 in Kuala Lumpur.

Durand-Gasselin T., Blangy L., Picasso C., de Franqueville H., Breton F., Amblard P., Cochard B., Louise C. et Nouy B., 2010. Sélection du palmier à huile pour une huile de palme durable et responsabilité sociale. OCL, 17(6): 385‑392.

Durand-Gasselin T., Kouame Kouame R., Cochard B., Adon B. et Amblard P., 2000. Diffusion variétale du palmier à huile (*Elaeis guineensis* Jacq.). Oléagineux, Corps Gras, Lipides, 7(2): 207‑214.

Dussert S., Guerin C., Andersson M., Joët T., Tranbarger T.J. *et al.*, 2013. Comparative Transcriptome Analysis of Three Oil Palm Fruit and Seed Tissues That Differ in Oil Content and Fatty Acid Composition. Plant Physiology, 162(3): 1337‑1358.

Eding H. et Meuwissen T.H.E., 2001. Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics, 118(3): 141‑159.

Eggen A., 2012. The development and application of genomic selection as a new breeding paradigm. Animal Frontiers, 2(1): 10‑15.

Elshire R.J., Glaubitz J.C., Sun Q., Poland J.A., Kawamoto K., Buckler E.S. et Mitchell S.E., 2011. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE, 6(5): e19379.

Elzo M.A., 1990. Recursive procedures to compute the inverse ofmultiple trait additive genetic covariance matrix in inbreed and non inbreed multibreed populations. J. Anim. Sci., 68: 1215–1228.

Emik L.O. et Terrill C.E., 1949. Systematic procedures for calculating inbreeding coefficients. The Journal of Heredity, 40(2): 51‑55.

Falconer D. et Mackay T., 1996. *Introduction to quantitative genetics*. Longman, Harlow, Essex, UK, 464 p.

FAO, 2009. How to feed the world in 2050 ? Food and Agriculture Organization of the United Nations, 35 p.

Fernández J. et Toro M.A., 2006. A new method to estimate relatedness from molecular markers. Molecular Ecology, 15(6): 1657‑1667.

Fisher R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh, 52: 399‑433.

Flint J. et Mackay T.F.C., 2009. Genetic architecture of quantitative traits in mice, flies, and humans. Genome Research, 19(5): 723‑733.

Fonds français pour l'alimentation et la santé, 2012. L'huile de palme : aspects nutritionnels, sociaux et environnementaux. Etat des lieux, 20 p.

Forni S., Aguilar I. et Misztal I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution, 43(1): 1.

Gallais A., 1990. *Théorie de la sélection en amélioration des plantes*. Masson, Collection Sciences agronomiques, 588 p.

Gallais A., 2009. *Hétérosis et variétés hybrides en amélioration des plantes*. Quae éditions, Synthèses, 376 p.

Gao H., Lund M.S., Zhang Y. et Su G., 2013. Accuracy of genomic prediction using different models and response variables in the Nordic Red cattle population. Journal of Animal Breeding and Genetics, 130(5): 333‑340.

García-Cortés L. et Toro M., 2006. Multibreed analysis by splitting the breeding values. Genetics Selection Evolution, 38(6): 601 ‑ 615.

Garrick D., Taylor J. et Fernando R., 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genetics Selection Evolution, 41(1): 55.

Gascon J.P. et de Berchoux C., 1964. Caractéristique de la production d'*Elaeis guineensis* (Jacq.) de diverses origines et de leurs croisements - Application à la sélection du palmier à huile. Oléagineux, 19(2): 75‑84.

Gascon J.P., Noiret J.M. et Bénard G., 1966. Contribution à l'étude de l'hérédité de la production de régimes d'*Elaeis guineensis* Jacq. - Application à la sélection du palmier à huile. Oléagineux, 21(11): 657‑661.

Gianola D., de los Campos G., Hill W.G., Manfredi E. et Fernando R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics, 183(1): 347‑363.

Gianola D. et van Kaam J.B.C.H.M., 2008. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics, 178(4): 2289‑2303.

Gilmour A.R., Gogel B.J., Cullis B.R. et Thompson R., 2009. *ASReml user guide release 3.0*. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK, www.vsni.co.uk, .

Goddard M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica, 136(2): 245‑257.

González-Camacho J.M., de los Campos G., Pérez P., Gianola D., Cairns J.E., Mahuku G., Babu R. et Crossa J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. Theoretical and Applied Genetics, 125(4): 759‑771.

Goudet J., 2013. *Hierfstat: estimation and tests of hierarchical F-statistics, R package version 0.04-10*.

Gowda M., Zhao Y., Wurschum T., Longin C.F., Miedaner T. *et al.*, 2014. Relatedness severely impacts accuracy of marker-assisted selection for disease resistance in hybrid wheat. Heredity, 112(5): 552‑561.

Grattapaglia D., 2014. Breeding forest trees by genomic selection: current progress and the way forward, pp. 651‑682 in *Genomics of Plant Genetic Resources*, Tuberosa R, Graner A, Frison E.

Guo G., Zhao F., Wang Y., Zhang Y., Du L. et Su G., 2014. Comparison of single-trait and multiple-trait genomic prediction models. BMC Genetics, 15(1): 30.

Gupta P.K., Rustgi S. et Kulwal P.L., 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Molecular Biology, 57(4): 461‑485.

Gutiérrez J.P., Cervantes I. et Goyache F., 2009. Improving the estimation of realized effective population sizes in farm animals. Journal of Animal Breeding and Genetics, 126(4): 327‑332.

Gutiérrez J.P., Cervantes I., Molina A., Valera M. et Goyache F., 2008. Individual increase in inbreeding allows estimating effective sizes from pedigrees. Genetics Selection Evolution, 40(4): 359 ‑ 378.

Gutiérrez J.P. et Goyache F., 2005. A note on ENDOG: a computer program for analysing pedigree information. Journal of Animal Breeding and Genetics, 122(3): 172‑176.

Habier D., Fernando R., Kizilkaya K. et Garrick D., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics, 12(1): 186.

Habier D., Fernando R.L. et Dekkers J.C.M., 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177(4): 2389‑2397.

Habier D., Fernando R.L. et Garrick D.J., 2013. Genomic BLUP Decoded: A Look into the Black Box of Genomic Prediction. Genetics, 194(3): 597‑607.

Habier D., Tetens J., Seefried F.-R., Lichtner P. et Thaller G., 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution, 42(1): 5.

Hardon J.J., 1970. Inbreeding in populations of the oil palm (*Elaeis guineensis* Jacq.) and its effect on selection. Oléagineux, 25: 449‑456.

Hardon J.J. et Thomas R.L., 1968. Breeding and selection of the oil palm in Malaya. Oléagineux, 23: 85–90.

Hayashi T. et Iwata H., 2013. A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. BMC Bioinformatics, 14(1): 34.

Hayes B. et Goddard M., 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution, 33(3): 209 - 229.

Hayes B.J., Visscher P.M., McPartlan H.C. et Goddard M.E., 2003. Novel Multilocus Measure of Linkage Disequilibrium to Estimate Past Effective Population Size. Genome Research, 13(4): 635-643.

Heffner E.L., Sorrells M.E. et Jannink J.-L., 2009. Genomic selection for crop improvement. Crop Sci., 49(1): 1-12.

Henderson C.R., 1950. Estimation of genetic parameters. Ann. Math. Statist, 21: 309-310.

Henderson C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31(2): 423-447.

Henderson C.R., 1984. *Applications of linear models in animal breeding*. University of Guelph, .

Henderson C.R., 1986. Statistical methods in animal improvement: Historical Overview, pp. 2-14 in *Advances in statistical methods for genetic improvement of lifestock*, Springer, Gianola D, Hammond K, Berlin.

Heslot N., Akdemir D., Sorrells M. et Jannink J.-L., 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and Applied Genetics, 127(2): 463-480.

Heslot N., Yang H.-P., Sorrells M.E. et Jannink J.-L., 2012. Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci., 52(1): 146-160.

Hill W.G., 1981. Estimation of effective population size from data on linkage disequilibrium. Genetical Research, 38: 209-216.

Hill W.G., 2010. Understanding and using quantitative genetic variation. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1537): 73-85.

Ho C.-L., Kwan Y.-Y., Choi M.-C., Tee S.-S., Ng W.-H. *et al.*, 2007. Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). BMC Genomics, 8(1): 381.

Howard R., Carriquiry A.L. et Beavis W.D., 2014. Parametric and Nonparametric Statistical Methods for Genomic Selection of Traits with Additive and Epistatic Genetic Architectures. G3: Genes|Genomes|Genetics, 4(6): 1027-1046.

Ibánez-Escriche N., Fernando R., Toosi A. et Dekkers J., 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution, 41(1): 12.

Isik F., 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forests, 45(3): 379‑401.

Jacquemard J.-C., Meunier J. et Bonnot F., 1981. Etude génétique de la reproduction d'un croisement chez le palmier à huile *Elaeis guineensis*, application à la production de semences sélectionnées et à l'amélioration. Oléagineux, 36(7): 343‑349.

Jannink J.-L., Lorenz A.J. et Iwata H., 2010. Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics, 9(2): 166‑177.

Jarquín D., Crossa J., Lacaze X., Du Cheyron P., Daucourt J. *et al.*, 2014. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theoretical and Applied Genetics, 127(3): 595‑607.

Jarquín D., Kocak K., Posadas L., Hyma K., Jedlicka J., Graef G. et Lorenz A., 2014. Genotyping by sequencing for genomic prediction in a soybean breeding population. BMC Genomics, 15(1): 740.

Jeennor S. et Volkaert H., 2013. Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). Tree Genetics & Genomes, .

Jia Y. et Jannink J.-L., 2012. Multiple trait genomic selection methods increase genetic value prediction accuracy. Genetics, .

Jonas E. et de Koning D.-J., 2013. Does genomic selection have a future in plant breeding? Trends in Biotechnology, 31(9): 497‑504.

Keenan K., McGinnity P., Cross T.F., Crozier W.W. et Prodöhl P.A., 2013. DiveRsity: an R package for the estimation and exploration of population genetics parameters and their associated errors. Methods in Ecology and Evolution, 4(8): 782‑788.

Kinghorn B.P., Hickey J.M. et Van Der Werf J.H.J., 2010. Reciprocal recurrent genomic selection for total genetic merit in crossbred individuals, pp. 36 in *Proceedings of the 9th World Congress on Genetics Applied to Livestock Production*, Leipzig, Germany.

Kumar S., Chagné D., Bink M.C.A.M., Volz R.K., Whitworth C. et Carlisle C., 2012. Genomic selection for fruit quality traits in apple (Malus×domestica Borkh.). PLoS ONE, 7(5): e36674.

Lande R. et Thompson R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124(3): 743‑756.

Laurie C.C., Chasalow S.D., LeDeaux J.R., McCarroll R., Bush D. *et al.*, 2004. The Genetic Architecture of Response to Long-Term Artificial Selection for Oil Concentration in the Maize Kernel. Genetics, 168(4): 2141‑2155.

Legarra A., Aguilar I. et Misztal I., 2009. A relationship matrix including full pedigree and genomic information. Journal of Dairy Science, 92(9): 4656‑4663.

Li C.C., Weeks D.E. et Chakravarti A., 1993. Similarity of DNA fingerprints due to chance and relatedness. Hum. Hered., 43(1): 45‑52.

Lo L.L., Fernando R.L. et Grossman M., 1993. Covariance between relatives in multibreed populations: additive model. Theoretical and Applied Genetics, 87(4): 423‑430.

Lo L.L., Fernando R.L. et Grossman M., 1997. Genetic evaluation by BLUP in two-breed terminal crossbreeding systems under dominance. Journal of Animal Science, 75(11): 2877‑2884.

Lorenz A.J., Chao S., Asoro F.G., Heffner E.L., Hayashi T., Iwata H., Smith K.P., Sorrells M.E. et Jannink J.-L., 2011. Genomic Selection in Plant Breeding: Knowledge and Prospects, pp. 77‑123 in *Advances in Agronomy*, édité par Donald L. Sparks. Academic Press.

Luyindula N., Mantantu N., Dumortier F. et Corley R.H.V., 2005. Effects of inbreeding on growth and yield of oil palm. Euphytica, 143(1-2): 9‑17.

Ly D., Hamblin M., Rabbi I., Melaku G., Bakare M. *et al.*, 2013. Relatedness and Genotype × Environment Interaction Affect Prediction Accuracies in Genomic Selection: A Study in Cassava. Crop Sci., 53(4): 1312‑1325.

Lynch M., 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol., 5(5): 584‑599.

Lynch M. et Walsh B., 1998. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA, 980 p.

Mackay I. et Powell W., 2007. Methods for linkage disequilibrium mapping in crops. Trends in Plant Science, 12(2): 57‑63.

Maenhout S., De Baets B. et Haesaert G., 2009. Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. Theoretical and Applied Genetics, 118(6): 1181‑1192.

Makgahlela M.L., Mäntysaari E.A., Strandén I., Koivula M., Nielsen U.S., Sillanpää M.J. et Juga J., 2013. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. Journal of Animal Breeding and Genetics, 130(1): 10‑19.

Malécot G., 1948. *Les mathématiques de l'hérédité*. Masson & Cie, Paris, 64 p.

Marchal A., 2014. Sélection génomique multivariée chez le palmier à huile. CIRAD, Rapport de stage de Master 2, Université Montpellier 2, 60 p.

Massman J., Gordillo A., Lorenzana R. et Bernardo R., 2013. Genomewide predictions from maize single-cross data. Theoretical and Applied Genetics, 126(1): 13‑22.

Meunier J. et Gascon J., 1972. Le schéma général d'amélioration du palmier à huile à l'IRHO. Oléagineux, 27(1): 1‑12.

Meunier J., Gascon J.P. et Noiret J.M., 1970. Hérédité des caractéristiques du régime d'*Elaeis guineensis* Jacq. en Côte d'Ivoire. Oléagineux, 25: 377‑382.

Meuwissen T., 2009. Accuracy of breeding values of « unrelated » individuals predicted by dense SNP genotyping. Genetics Selection Evolution, 41(1): 35.

Meuwissen T.H., 1997. Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science, 75(4): 934‑940.

Meuwissen T.H.E., Hayes B.J. et Goddard M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4): 1819‑1829.

Mrode R.A., 2005. *Linear models for the prediction of animal breeding values*. CABI, Oxfordshire, UK, 344 p.

Muir W.M., 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. Journal of Animal Breeding and Genetics, 124(6): 342‑355.

Muranty H., Jorge V., Bastien C., Lepoittevin C., Bouffier L. et Sanchez L., 2014. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. Tree Genetics & Genomes, 1‑20.

Murphy D.J., 2014. The Future of Oil Palm as a Major Global Crop: Opportunities and Challenges. Journal of Oil Palm Research, 26(1): 1‑24.

Myles S., 2013. Improving fruit and wine: what does genomics have to offer? Trends in Genetics, 29(4): 190‑196.

Neves H.H., Carvalheiro R. et Queiroz S., 2012. A comparison of statistical methods for genomic selection in a mice population. BMC Genetics, 13(1): 100.

Noiret J.M., Gascon J.P. et Bénard G., 1966. Contribution a l'étude de l'hérédité des caractéristiques de la qualité du régime et du fruit d'*Elaeis guineensis* Jacq. - Application à la sélection du palmier à huile. Oléagineux, 21(6): 343‑349.

Obisesan I.O. et Fatunla T., 1983. Genotype x environment interaction for bunch yield and its components in the oil palm (*Elaeis guineensis*, Jacq.). Theoretical and Applied Genetics, 64(2): 133‑136.

Ooi S.C., Hardon J.J. et Phang S., 1973. Variability in the Deli dura breeding population of the oil palm (*Elaeis guineensis* Jacq.). I. Components of bunch yield. Malay. agric. J., 49: 112‑121.

Ornella L., Perez P., Tapia E., Gonzalez-Camacho J.M., Burgueno J. *et al.*, 2014. Genomic-enabled prediction with classification algorithms. Heredity, 112(6): 616‑626.

Ostersen T., Christensen O., Henryon M., Nielsen B., Su G. et Madsen P., 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. Genetics Selection Evolution, 43(1): 38.

Park T. et Casella G., 2008. The Bayesian Lasso. Journal of the American Statistical Association, 103(482): 681‑686.

Patry C. et Ducrocq V., 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of Dairy Science, 94(2): 1011‑1020.

Pérez P. et de los Campos G., 2013. *BGLR: A Statistical Package for Whole Genome Regression and Prediction*. R package version 1.0.2, .

Pérez P., de los Campos G., Crossa J. et Gianola D., 2010. Genomic-Enabled Prediction Based on Molecular Markers and Pedigree Using the Bayesian Linear Regression Package in R. Plant Gen., 3(2): 106‑116.

Piepho H.P., Möhring J., Melchinger A.E. et Büchse A., 2008. BLUP for phenotypic selection in plant breeding and variety testing. Euphytica, 161(1-2): 209‑228.

Poland J.A. et Rife T.W., 2012. Genotyping-by-Sequencing for Plant Breeding and Genetics. Plant Gen., 5(3): 92‑102.

Poland J., Endelman J., Dawson J., Rutkoski J., Wu S. *et al.*, 2012. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Gen., 5(3): 103‑113.

Pszczola M., Strabel T., Mulder H.A. et Calus M.P.L., 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. Journal of Dairy Science, 95(1): 389‑400.

Purba A.R., Flori A., Baudouin L. et Hamon S., 2001. Prediction of oil palm (*Elaeis guineensis* Jacq.) agronomic performances using the best linear unbiased predictor (BLUP). Theoretical and Applied Genetics, 102(5): 787‑792.

Rafii M.Y., Jalani B.S., Rajanaidu N., Kushairi A., Puteh A. et Latif M.A., 2012. Stability analysis of oil yield in oil palm (*Elaeis guineensis*) progenies in different environments. Genet. Mol. Res., 11(4): 3629‑3641.

Rance K.A., Mayes S., Price Z., Jack P.L. et Corley R.H.V., 2001. Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 103(8): 1302‑1310.

R Core Team, 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, .

Reif J.C., Zhao Y., Würschum T., Gowda M. et Hahn V., 2013. Genomic prediction of sunflower hybrid performance. Plant Breeding, 132(1): 107‑114.

Resende M.D.V., Resende M.F.R., Sansaloni C.P., Petroli C.D., Missiaggia A.A. *et al.*, 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing

heritability and accelerating breeding for complex traits in forest trees. New Phytologist, 194(1): 116‑128.

Riedelsheimer C., Czedik-Eysenberg A., Grieder C., Lisec J., Technow F. *et al.*, 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet, 44(2): 217‑220.

Rincent R., Laloë D., Nicolas S., Altmann T., Brunel D. *et al.*, 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). Genetics, .

De Roos A.P.W., Hayes B.J. et Goddard M.E., 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics, 183(4): 1545‑1553.

Russell J. et Fewster R., 2009. Evaluation of the Linkage Disequilibrium Method for Estimating Effective Population Size, pp. 291‑320 in *Modeling Demographic Processes In Marked Populations*, édité par D. Thomson, E. Cooch, et M. Conroy. Environmental and Ecological Statistics, Springer US.

Rutkoski J.E., Poland J.A., Singh R.P., Huerta-Espino J., Bhavani S., Barbier H., Rouse M.N., Jannink J.-L. et Sorrells M.E., 2014. Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. Plant Gen., 0(0): -.

Saatchi M., McClure M., McKay S., Rolf M., Kim J. *et al.*, 2011. Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. Genetics Selection Evolution, 43(1): 40.

Sánchez L., Toro M.A. et García C., 1999. Improving the Efficiency of Artificial Selection: More Selection Pressure With Less Inbreeding. Genetics, 151(3): 1103‑1114.

Sánchez L., Yanchuk A. et King J., 2008. Gametic models for multitrait selection schemes to study variance of response and drift under adverse genetic correlations. Tree Genetics & Genomes, 4(2): 201‑212.

Sanger F. et Coulson A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 94(3): 441‑448.

Sanger F., Nicklen S. et Coulson A.R., 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12): 5463‑5467.

Schadt E.E., Turner S. et Kasarskis A., 2010. A window into third-generation sequencing. Human Molecular Genetics, 19(R2): R227‑R240.

Schefers J.M. et Weigel K.A., 2012. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers, 2(1): 4‑9.

Schnell F.W. et Cockerham C.C., 1992. Multiplicative vs. arbitrary gene action in heterosis. Genetics, 131(2): 461‑469.

Schön C.C., Utz H.F., Groh S., Truberg B., Openshaw S. et Melchinger A.E., 2004. Quantitative Trait Locus Mapping Based on Resampling in a Vast Maize Testcross Experiment and Its Relevance to Quantitative Genetics for Complex Traits. Genetics, 167(1): 485‑498.

Seng T.-Y., Mohamed Saad S.H., Chin C.-W., Ting N.-C., Harminder Singh R.S., Qamaruz Zaman F., Tan S.-G. et Syed Alwee S.S.R., 2011. Genetic Linkage map of a high yielding FELDA Deli×Yangambi oil palm cross. PLoS ONE, 6(11): e26593.

Singh R., Ong-Abdullah M., Low E.-T.L., Manaf M.A.A., Rosli R. *et al.*, 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature, 500(7462): 335‑339.

Slatkin M., 2008. Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nat Rev Genet, 9(6): 477‑485.

Soh A.C., 1994. Ranking parents by best linear unbiased prediction (BLUP) breeding values in oil palm. Euphytica, 76(1-2): 13‑21.

Solberg T.R., Sonesson A.K., Woolliams J.A. et Meuwissen T.H.E., 2008. Genomic selection using different marker types and densities. Journal of Animal Science, 86(10): 2447‑2454.

Sonesson A., Woolliams J. et Meuwissen T., 2012. Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution, 44(1): 27.

Souchard V., 2013. Sélection du palmier à huile pour la régularité de la production de régimes. CIRAD, Rapport de stage de Master 2, Université Montpellier 2, 31 p.

De Souza Jr C.L., 1992. Interpopulation genetic variances and hybrid breeding programs. Brazilian Journal of Genetics, 15(3): 643‑656.

Speed D. et Balding D.J., 2014. MultiBLUP: improved SNP-based prediction for complex traits. Genome Research, .

Strandén I. et Mäntysaari E.A., 2013. Use of random regression model as an alternative for multibreed relationship matrix. Journal of Animal Breeding and Genetics, 130(1): 4‑9.

Stuber C.W. et Cockerham C.C., 1966. Gene effects and variances in hybrid populations. Genetics, 54(6): 1279‑1286.

Su G., Christensen O.F., Ostersen T., Henryon M. et Lund M.S., 2012. Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS ONE, 7(9): e45293.

Sun C., VanRaden P.M., Cole J.B. et O'Connell J.R., 2014. Improvement of Prediction Ability for Genomic Selection of Dairy Cattle by Including Dominance Effects. PLoS ONE, 9(8): e103934.

Sun X., Peng T. et Mumm R.H., 2011. The role and basics of computer simulation in support of critical decisions in plant breeding. Molecular Breeding, 28(4): 421‑436.

Sved J.A., 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite population. Theoret. Pop. Biol., 2: 125–141.

Technow F., Riedelsheimer C., Schrag T. et Melchinger A., 2012. Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. Theoretical and Applied Genetics, 125(6): 1181‑1194.

Tee S.-S., Tan Y.-C., Abdullah F., Ong-Abdullah M. et Ho C.-L., 2013. Transcriptome of oil palm (*Elaeis guineensis* Jacq.) roots treated with Ganoderma boninense. Tree Genetics & Genomes, 9(2): 377‑386.

Thomas R.L., Watson I. et Hardon J.J., 1969. Inheritance of some components of yield in the 'Deli dura variety' of oil palm. Euphytica, 18: 92‑100.

Thomsen H., Reinsch N., Xu N., Looft C., Grupe S. *et al.*, 2001. Comparison of estimated breeding values, daughter yield deviations and de-regressed proofs within a whole genome scan for QTL. Journal of Animal Breeding and Genetics, 118(6): 357‑370.

Tibshirani R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267‑288.

Ting N.-C., Jansen J., Mayes S., Massawe F., Sambanthamurthi R. *et al.*, 2014. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. BMC Genomics, 15(1): 309.

Toosi A., Fernando R.L. et Dekkers J.C.M., 2010. Genomic selection in admixed and crossbred populations. Journal of Animal Science, 88(1): 32‑46.

Toro M. et Perez-Enciso M., 1990. Optimization of selection response under restricted inbreeding. Genetics Selection Evolution, 22(1): 93 ‑ 107.

Tranbarger T.J., Dussert S., Joët T., Argout X., Summo M. *et al.*, 2011. Regulatory Mechanisms Underlying Oil Palm Fruit Mesocarp Maturation, Ripening, and Functional Specialization in Lipid and Carotenoid Metabolism. Plant Physiology, 156(2): 564‑584.

Tranbarger T., Kluabmongkol W., Sangsrakru D., Morcillo F., Tregear J., Tragoonrung S. et Billotte N., 2012. SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. BMC Plant Biology, 12(1): 1.

Ukoskit K., Chanroj V., Bhusudsawang G., Pipatchartlearnwong K., Tangphatsornruang S. et Tragoonrung S., 2014. Oil palm (*Elaeis guineensis* Jacq.) linkage map, and quantitative trait locus analysis for sex ratio and related traits. Molecular Breeding, 33(2): 415‑424.

USDA, 2014. Oilseeds: world market and trade. Foreign Agricultural Service, Circular Series May 2014.

VanRaden P.M., 2007. Genomic measures of relationship and inbreeding. Interbull Bulletin, 37: 33‑36.

Verrier E., Brabant P. et Gallais A., 2001. *Faits et concepts de base en génétique quantitative*.

Viana J.M.S., Valente S.F., Scapim C.A., de Resende M.D.V. et e Silva F.F., 2011. Genetic evaluation of tropical popcorn inbred lines using BLUP. Maydica, 56: 273‑281.

Villumsen T.M., Janss L. et Lund M.S., 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. Journal of Animal Breeding and Genetics, 126(1): 3‑13.

Vitezica Z.G., Varona L. et Legarra A., 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics, 195(4): 1223‑1230.

Walsh B., 2013.Lecture 28: BLUP and genomic selection, Bruce Walsh lecture notes, Synbreed course, version 11 July 2013, 55p. .

Wang C.-L., Ma P.-P., Zhang Z., Ding X.-D., Liu J.-F., Fu W.-X., Weng Z.-Q. et Zhang Q., 2012. Comparison of five methods for genomic breeding value estimation for the common dataset of the 15th QTL-MAS Workshop. BMC Proceedings, 6(Suppl 2): S13.

Wang J., 2005. Estimation of effective population sizes from data on genetic markers. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1459): 1395‑1409.

Wang J., 2014. Marker-based estimates of relatedness and inbreeding coefficients: an assessment of current methods. Journal of Evolutionary Biology, 27(3): 518‑530.

Waples R.S., 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conservation Genetics, 7(2): 167‑184.

Waples R.S. et Do C., 2008. LDNE: a program for estimating effective population size from data on linkage disequilibrium. Molecular Ecology Resources, 8(4): 753‑756.

Waples R.S. et Do C., 2010. Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. Evolutionary Applications, 3(3): 244‑262.

Weir B.S., 1979. Inferences about linkage disequilibrium. Biometrics, 35: 235‑254.

Weir B.S., 1996. *Genetic data analysis*. Sinauer Associates, Sunderland, MA, 445 p.

Weir B.S. et Cockerham C.C., 1984. Estimating F-Statistics for the analysis of population structure. Evolution, 38: 1358‑1370.

Wong C.K. et Bernardo R., 2008. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics, 116(6): 815‑824.

Wright S., 1922. Coefficients of inbreeding and relationship. Amer. Nat., 56: 330–338.

Wright S., 1931. Evolution in Mendelian populations. Genetics, 16(2): 97‑159.

Wu H.X. et Sánchez L., 2011. Effect of selection method on genetic correlation and gain in a two-trait selection scheme. Australian Forestry, 74(1): 36‑42.

Xu S., Zhu D. et Zhang Q., 2014. Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proceedings of the National Academy of Sciences, 111(34): 12456‑12461.

Zapata-Valenzuela J., Isik F., Maltecca C., Wegrzyn J., Neale D., McKeand S. et Whetten R., 2012. SNP markers trace familial linkages in a cloned population of *Pinus taeda*— prospects for genomic selection. Tree Genetics & Genomes, 8(6): 1307‑1318.

Zeng J., Toosi A., Fernando R., Dekkers J. et Garrick D., 2013. Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genetics Selection Evolution, 45(1): 11.

Zhang Z., Liu J., Ding X., Bijma P., de Koning D.-J. et Zhang Q., 2010. Best Linear Unbiased Prediction of Genomic Breeding Values Using a Trait-Specific Marker-Derived Relationship Matrix. PLoS ONE, 5(9): e12648.

Zhang Z., Ober U., Erbe M., Zhang H., Gao N., He J., Li J. et Simianer H., 2014. Improving the Accuracy of Whole Genome Prediction for Complex Traits Using the Results of Genome Wide Association Studies. PLoS ONE, 9(3): e93017.

Zhao Y., Zeng J., Fernando R. et Reif J.C., 2013. Genomic Prediction of Hybrid Wheat Performance. Crop Sci., 53(3): 802‑810.

Annexes

Annexe 1 : Le palmier à huile (*Elaeis guineensis* Jacq) et la production d'huile

A

http://commons.wikimedia.org

B

http://www.nafas.com.my

C

A Plantation commerciale (Malaisie)
B Récolte (Malaisie)
C Récolte (Bénin)
D Inflorescence mâle à maturité
E Régime mûr
F Fruits mûrs

D

E

F

G

H

http://commons.wikimedia.org

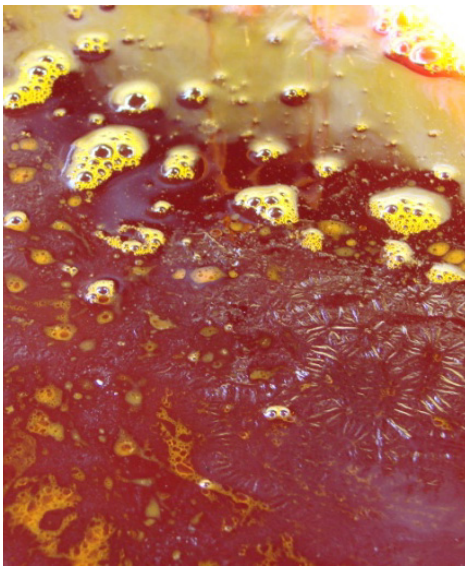
I

G Coupe transversale de fruit tenera (type commercial)
H Usine et régimes (Côte d'Ivoire)
I Plantation et usine (Indonésie)
J Presse semi-industrielle
K Presse artisanale
L Huile rouge (huile de palme brute, non raffinée)

J

K

L



Annexe 2 : L'amélioration génétique et la production de semences

M

M Palmier à huile élite de la population Deli (PO3600D, Pobè, Bénin)
N Inflorescence mâle dégagée avant ensachage en prévision de la récolte du pollen
O Inflorescence femelle ensachée, à maturité pour la fécondation artificielle
P Fécondation artificielle
Q Couronne de palmier à huile de la population Deli chargée de régimes de fécondations artificielles
R Régime de fécondation artificielle proche de la maturité (Deli) et sélectionneur (PalmElit)
S Graines sèches
T Graines germées
U Coupe longitudinale de fruit dura
V Coupe transversale de fruit pisifera
W Composantes du régime (pédoncule, épillets, fruits et graines)
X Décompte et pesée des régimes
Y Mesure de hauteur
Z Extracteurs de Soxhlet destinés à mesurer le pourcentage d'huile dans la pulpe

N

O

P

Q

R

S

T

U

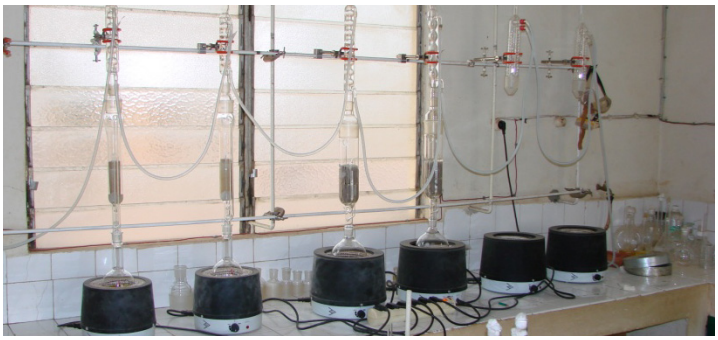
V

W

X

Y

Z



Annexe 3 : Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population

Cros D., Sánchez L., Cochard B., Samper P., Denis M., Bouvet J.-M. et Fernández J., 2014. Theoretical and Applied Genetics, 127(4): 981‑994.

Theor Appl Genet (2014) 127:981–994
DOI 10.1007/s00122-014-2273-3

ORIGINAL PAPER

Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population

David Cros · Leopoldo Sánchez · Benoit Cochard · Patrick Samper · Marie Denis · Jean-Marc Bouvet · Jesús Fernández

Received: 21 March 2013 / Accepted: 22 January 2014 / Published online: 7 February 2014

Abstract
Key message **Explicit pedigree reconstruction by simulated annealing gave reliable estimates of genealogical coancestry in plant species, especially when selfing rate was lower than 0.6, using a realistic number of markers.** Genealogical coancestry information is crucial in plant breeding to estimate genetic parameters and breeding values. The approach of Fernández and Toro (Mol Ecol 15:1657–1667, 2006) to estimate genealogical coancestries from molecular data through pedigree reconstruction was limited to species with separate sexes. In this study it was extended to plants, allowing hermaphroditism and monoecy, with possible selfing. Moreover, some improvements were made to take previous knowledge on the population demographic history into account. The new method was validated using simulated and real datasets. Simulations showed that accuracy of estimates was high with 30 microsatellites, with the best results obtained for selfing rates below 0.6. In these conditions, the root mean square error (RMSE) between the true and estimated genealogical coancestry was small (<0.07), although the number of ancestors was overestimated and the selfing rate could be biased. Simulations also showed that linkage disequilibrium between markers and departure from the Hardy–Weinberg equilibrium in the founder population did not affect the efficiency of the method. Real oil palm data confirmed the simulation results, with a high correlation between the true and estimated genealogical coancestry (>0.9) and a low RMSE (<0.08) using 38 markers. The method was applied to the Deli oil palm population for which pedigree data were scarce. The estimated genealogical coancestries were highly correlated (>0.9) with the molecular coancestries using 100 markers. Reconstructed pedigrees were used to estimate effective population sizes. In conclusion, this method gave reliable genealogical coancestry estimates. The strategy was implemented in the software MOLCOANC 3.0.

Communicated by M. Frisch.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-014-2273-3) contains supplementary material, which is available to authorized users.

D. Cros (✉) · B. Cochard · P. Samper · M. Denis · J.-M. Bouvet
Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit (AGAP), CIRAD, International campus of Baillarguet, TA A-108/C, 34398 Montpellier Cedex 5, France
e-mail: david.cros@cirad.fr

L. Sánchez
Forest Tree Improvement, Genetics and Physiology Research Unit (AGPF), INRA, 2163 Avenue de la Pomme de Pin, CS 40001 Ardon, 45075 Orleans Cedex 2, France

J. Fernández
Departamento de Mejora Genética Animal, INIA, Ctra. Coruña Km 7.5, 28040 Madrid, Spain

Introduction

Knowledge of genetic relationships between individuals in plant populations is of major interest for breeding as it allows estimation of the genetic parameters (heritabilities and genetic correlations), breeding values and effective population sizes. Moreover, this information is needed to reduce the increase of inbreeding and loss of diversity through the control of global coancestry. These values are conventionally obtained from genealogical information.

However, pedigrees of breeding populations of plant species may be completely, or at least partially, unknown

for various reasons. Breeding of many plant species includes uncontrolled open pollination or controlled pollination with a pollen mixture, for instance to carry out open pollinated or polycross progeny tests, as in forage crops and forest trees. Missing genealogical information could also be due to the existence of cutoff dates for recording the pedigree (i.e., pedigree recording does not start from the unrelated founder individuals) and long breeding cycles. Furthermore, breeding programs include many error-prone steps, including pollination, seed preparation, germination, planting, etc. Mistakes may occur even for controlled crosses between clearly identified parents, leading to illegitimate progenies (one or both parents actually unknown) or contamination (some illegitimate individuals in the progeny). This has been reported many times, for instance in sugarcane (McIntyre and Jackson 2001), oil palm (Corley 2005), Douglas-fir and loblolly pine (Adams et al. 1988). In addition, individuals of unknown origin can enter the breeding population, such as when coming from another breeding program.

Completely missing genealogical information, small depth pedigrees or erroneous ones have negative effects on breeding programs, often due to erroneous estimations of genetic parameters and breeding values. In diallel progeny trials, Ericsson (1999) showed that 0.5 % of trees with misidentified pedigree were enough to downward bias additive variance and heritability estimates. Atkin et al. (2009) found that the accuracy of additive variance components and breeding values improved when more pedigree information was used in analyses. In polycross progeny tests, the unknown relative contribution of male parents can be detrimental (Kumar et al. 2007) and reconstructing male parentage increased the genetic gain (Doerksen and Herbinger 2010). Furthermore, assuming that individuals newly imported into a breeding population are unrelated to the main population may lead to erroneous management decisions if a common ancestral origin actually exists.

Therefore, there is substantial interest in having good knowledge of the pedigree of breeding populations of plant species. In particular, this would result in reliable estimates of the genealogical coancestry and, consequently, in better breeding management. Fernández and Toro (2006) proposed an approach (FT method, thereafter) for estimating the genealogical coancestry between contemporaneous genotyped individuals through the construction of an explicit virtual genealogy. Through a *simulated annealing* algorithm (Kirkpatrick et al. 1983), FT finds the pedigree that yields a genealogical coancestry matrix with the highest correlation relative to the actual molecular coancestry matrix, or any other provided relatedness matrix. In other words, FT matches molecular coancestries within a given current population to the most likely compatible pedigree. Several parameters help to reduce the space of parametric solutions and to find a solution closer to the true pedigree, like the number of previous discrete generations to reconstruct, the maximum number of sires and dams in those previous generations and, if available, a known part of the pedigree. The advantages of FT over other methods (see Blouin 2003; Butler et al. 2004; Pemberton 2008 for reviews) are that it does not require knowledge of the true allelic frequencies in the base population (i.e., founder individuals from which the genotyped individuals derived) and that Hardy–Weinberg and linkage equilibria in the base population are not necessary. Moreover, it can manage any degree of complexity in relatedness between individuals and always provides congruent relationships, resulting in positive definite pedigree-based coancestry matrices.

The goals of this study were first to extend the capabilities of the FT method, in order to make it suitable for plant species, and second to demonstrate these new capabilities with simulated and true plant breeding populations.

FT was developed for dioecious species, i.e., species with separate sexes, especially animals. Consequently, any virtual ancestor acting as a node in the simulated pedigree in FT was either a male or a female, never both simultaneously. Therefore selfing, which is possible in many plant species and a tool for their breeding, was not possible. We extended the FT approach to encompass monoecy and hermaphroditism, with the possibility of selfing (i.e., mixed-mating, where a fraction of the progeny is derived from self-fertilization and the remainder from outcrossing), either when mixed-mating is the natural mode of reproduction or because this is artificially forced for breeding purposes. The first generation with selfings is a user-defined parameter, thus accounting for the possibility of natural cross-fertilizing species that, at a particular time, entered a breeding program in which self-fertilization could be artificially conducted. Another limitation of the FT approach was that the size of the virtual population was constant over generations. In order to take knowledge of the population demographic history into account, we made FT able to consider a variable number of ancestors through generations. The possibility of starting from related founders was also implemented, as it helps to get solutions which fit better to the real data.

To demonstrate the capabilities of the new method (FT*, thereafter), we used simulated data of a mixed-mating species and real data from two oil palm (*Elaeis guineensis*) breeding populations. Regarding the simulated data, we evaluated the effects of different selfing rates, percentages of unknown parentages and numbers of markers on the accuracy of the method, as well as the effects of departures from the ideal situation of Hardy–Weinberg and linkage equilibria. The real data involved the Yangambi (Africa) and the Deli (Asia) oil palm populations, for which important breeding efforts are currently under way. The Yangambi

population was used to validate FT*, as the pedigree is well known back to founder individuals. For the Deli population, the pedigree data are scarce and we used this population to illustrate one key application of pedigree reconstruction with FT*, the estimation of the pedigree-based effective population size (*Ne*) via the approach developed by Gutiérrez et al. (2008, 2009) and Cervantes et al. (2011). *Ne* is a parameter of interest in oil palm and no estimates are available. This species is "temporally dioecious", producing male and female inflorescences in an alternating cycle on the same plant, resulting in an allogamous mode of reproduction. Consequently, selfing does not occur in nature, but it has been used by breeders at some point in its pedigree. The inference of the extent of man-made selfing events is therefore of importance for current breeding population management. The *Ne* values obtained with the reconstructed pedigree were compared to those obtained via the method of Hill (1981) and Waples (2006) which is based only on linkage disequilibrium and independent of pedigree data.

Materials and methods

Original algorithm and new additions

The original method of Fernández and Toro (2006) starts with the generation of a random pedigree for the genotyped individuals. Alternative solutions are generated by randomly substituting one of the ancestors for another of the same generation. Alternative solutions are checked to avoid incompatible full-sib families at the molecular level. Valid solutions are used to calculate the genealogical coancestry matrix of genotyped individuals with the tabular method (Emik and Terrill 1949) and its correlation with the molecular coancestry matrix, which is calculated according to Eding and Meuwissen (2001). Coancestry coefficients are defined as the probability that two alleles taken at random, one from each individual, are identical by descent (genealogical coancestry) or by state (molecular coancestry). The probability of acceptance of alternative solutions is a function of the difference in the correlation of genealogical coancestry with molecular coancestry between the alternative and current solution and the cooling factor. The optimal solution is reached when no alternative solutions are accepted for 5,000 changes at a given 'temperature' or when the maximum number of steps is performed.

The FT method was modified to include the following new features:

1. Monoecy and hermaphroditism were implemented in addition to dioecy, so that the algorithm can create virtual ancestors either with separate sexes or with both sexes at once.
2. The possibility of selfing was implemented, as an option within monoecy and hermaphroditism, from the base population (i.e., in the whole pedigree) or from a later user-specified generation. This latter possibility is relevant when known artificial self-fertilization has been recently started in a species with non-natural selfing. When selfing is allowed, the self coancestries are also included in the calculation of the correlation between molecular and estimated genealogical coancestry matrices. Furthermore, a new rule was added for selfing when checking Mendelian inheritance in initial and alternative solutions considered by the simulated annealing algorithm, as no more than two alleles could exist in a full-sib family arising from selfing. Modifications 1. and 2. allow for the application of the method to the different modes of sexual reproduction existing in plant species.
3. The maximum number of ancestors per generation can be separately defined by the user. In this way, any previous information on the demographic history of the population can be taken into account (e.g., number of founders, known bottlenecks or expansions). These explicit limits in the size of generations reduce the space of feasible solutions for the optimization process. Thus the optimal solution is found more easily by the algorithm and is expected to be closer to the true pedigree.
4. We also included the possibility of accounting for a predefined coancestry matrix between founders, whenever real knowledge on the origin of the oldest known ancestors is available or simply to compare different hypotheses about the foundation of the population (i.e., from related or unrelated individuals).

All the other features of the FT method (coping with genotyping errors, including known relationships, etc.) were included in the new version too. The original as well as the new code were written in FORTRAN and a compiled version of the software for Windows platforms can be freely downloaded (MOLCOANC version 3.0, at http://dl.dropbox.com/u/5714008/Fernandez.htm). In addition, pedigrees are now automatically drawn with 'Pedigraph' (Garbe and Da 2008) if this software is already installed in the computer.

Testing the effect of selfing rate and marker numbers with simulated data

We simulated pedigrees of an expanding population of a hermaphroditic species, with different selfing rates and variable numbers of simple sequence repeat markers (SSR), in order to test the effects of these two parameters on the accuracy of genealogical coancestries estimated by FT*. The population started with five founders and reached 40 individuals in the sixth generation. In the founder

generation, 10, 30 or 90 SSR were simulated by randomly drawing without replacement alleles from a pool of 10 equiprobable alleles, independently for each marker. Consequently, the markers were expected to be in Hardy–Weinberg and linkage equilibria at the initial generation (base population). Notice that the sampling process produced data with a strong variation in allelic frequencies within loci and actual number of alleles between loci. This way simulated data better mimics the kind of scenarios which could be found in real data. Sequence repeat markers were evenly distributed along a genome of 10 chromosomes of 160 cM each, resulting in a genome length close to the 1.7 M found in oil palm (Billotte et al. 2005), in order to facilitate the comparison of results between simulations and real datasets. Individuals were considered as male and/or female and mated randomly, at first with the exclusion of selfing, while making sure that there was at least one mating per individual. Afterwards, six selfing rates were tested (0, 0.2, 0.4, 0.6, 0.8 and 1). The selfing rate was defined as the ratio of the number of individuals being the offspring of a self-fertilization to the total number of individuals (excluding founders). To achieve the defined selfing rates, a corresponding number of random crosses were converted into selfings. When the selfing rate differed from zero, selfing was allowed from the first generation and the same selfing rate was applied to each generation. Fifty pedigrees were simulated for each combination of selfing rate and number of SSR. Segregation of founder alleles in the pedigree was simulated with the R 'pedantics' package (Morrissey and Wilson 2010), with the mutation rate set at zero. FT* was applied to individuals of the last generation to reconstruct their pedigree from their molecular data and to estimate their genealogical coancestry.

For this and the following simulations, as well as for the real datasets, details about datasets and parameters used for the pedigree reconstruction algorithm are given in Table 1. For the *simulated annealing*, the maximum number of steps allowed was 150, the number of solutions tested at each step was 5,000, the initial temperature was 0.9 and the rate of temperature decrease was 0.99, in all situations.

Table 1 Marker data, true pedigree data and control parameters for simulated annealing for each case studied

Case study					
	Simulation			Oil palm real data	
	Selfing rate	% of unknown parentages	HW/LD	Yangambi	Deli
True pedigree data					
Number of generations	6	5	6	5	Unknown
Number of individuals per past generation[a]	5, 10, 15, 20, 30	5, 10, 15, 20	20, 25, 30, 40, 50	9, 7, 7, 5	Unknown
First generations with selfings allowed	1	1	1	2	Unknown
Marker data					
Number of SSR	10, 30, 90	10, 36, 63, 90	92 ± 6, 103 ± 8[b,e]	6–166	8–160
Number of genotyped individuals[c]	40	25	60	16	104
Average number of alleles per SSR	5.5 ± 1.0[b]	5.4 ± 1.0[b]	2.2 ± 0.4, 2.1 ± 0.4[b,e]	3.6 (2–6)[d]	2.7 (2–5)[4]
Control parameters for MOLCOANC					
Number of generations to reconstruct	5	4	5	4	7–9
Maximum number of individuals per reconstructed generation[c]	10, 20, 30, 40, 60	10, 20, 30, 40	40, 50, 60, 80, 100	64, 64, 64, 32	4, 30, 60, 31, 19, 3, 8–4, 30, 30, 30, 60, 31, 19, 3, 8
First generations with selfing allowed	1	1	1	2	4–5

HW Hardy–Weinberg disequilibrium, *LD* linkage disequilibrium

[a] Starting from the founder generation

[b] Computed over all replicates (±SD)

[c] i.e., Individuals of the last generation

[d] Range for the studied dataset

[e] Low and disequilibriums conditions, respectively

Testing the effect of percentage of unknown parentages and marker numbers with simulated data

These simulated datasets were used to investigate the effect of the percentage of unknown parentages and numbers of SSR on the accuracy of genealogical coancestries estimated by FT*. Pedigree and molecular data were simulated in the same way as in the previous simulation, except that the selfing rate was not controlled and therefore selfings occurred randomly. We simulated 11 datasets as replicates. Four levels of parentage removal from known lineages (25, 50, 75 and 100 % of the known parentages) and four numbers of SSR (10, 36, 63 and 90) were tested. For each percentage of unknown parentage we conducted eight repetitions, by randomly choosing the parentages to remove, and for each number of SSR we conducted eight repetitions, using the genotype for randomly chosen subsets of SSR. When removing known parentages, random individuals were selected in the pedigree and their two parents were set as missing. FT* was applied to the individuals of the last generation to reconstruct their pedigree from their molecular data and to estimate their genealogical coancestry.

Testing the effect of LD and HW in the base population with simulated data

The effect of linkage disequilibrium (LD) and departure from Hardy–Weinberg equilibrium (HW) in the base population on the accuracy of genealogical coancestries estimated by FT* was also assessed by simulation.

We first generated a base population departing from HW and with LD ('NON IDEAL' base population) by previously simulating 100 discrete generations with a constant generation size of 20 individuals. In the initial generation, 150 SSR were simulated by randomly drawing without replacement alleles from a pool of 15 equiprobable alleles, independently for each marker.

Sequence repeat markers were evenly distributed along a genome of 10 chromosomes of 150 cM each. Matings were done at random while imposing a minimum percentage of selfings of 20 % per generation, in order to further increase random genetic drift. Each individual had an equal contribution to the next generation. Marker allele segregation was simulated using 'pedantics', with the mutation rate set at zero. The last generation (100) constituted the 'NON IDEAL' base population. From this, we derived an 'IDEAL' population, by randomizing alleles present in the 'NON IDEAL' population within each SSR among the 20 individuals. In this way, the degree of polymorphism and the allelic frequencies remained the same for each marker between the two populations, while they differed by their LD and by the magnitude of the departure from HW equilibrium. LD was measured for each pair of SSR by the squared Pearson correlation (r^2) using the R 'gap' package (Zhao 2007). These values were used in a non-parametric paired test of Wilcoxon to check if the mean r^2 over all SSR pairs was significantly smaller in the 'IDEAL' base population than in the 'NON IDEAL' base population, i.e., that the simulation successfully created a significantly higher LD in the 'NON IDEAL' base population. Similarly, HW equilibrium was assessed for each SSR with the exact tests of Emigh (1980) for biallelic markers or Guo and Thompson (1992) for multiallelic markers. The p value of each test was kept, with p values lower than 0.05 indicating significant departure from HW. They were used in a non-parametric paired test of Wilcoxon to check if the mean p value over all SSR was significantly smaller in the 'NON IDEAL' base population than in the 'IDEAL' base population, i.e., that the simulation successfully created HW deviation in the 'NON IDEAL' base population. This process was replicated 75 times and only 24 replicates were finally used, as their p value was lower than 0.05 for the two Wilcoxon tests.

Once pairs of 'IDEAL' and 'NON IDEAL' base populations were successfully created, we simulated six additional generations of an expanding pedigree with a random mating regime, including the possibility of selfing. This pedigree was the same for a given pair of 'IDEAL' and 'NON IDEAL' base populations. FT* was applied to the 60 individuals of the last generation to reconstruct their pedigree relationships from the molecular data and to estimate their genealogical coancestry. Consequently, it was possible to compare the accuracy of the method when markers were at equilibrium (both HW and linkage) or when they departed from these 'ideal conditions'.

In order to test the effect of a stronger departure from ideal conditions, this procedure was repeated with a higher imposed minimum percentage of selfing when generating the 'NON IDEAL' base populations (40 %, instead of 20 %). To compensate for the higher random genetic drift associated with the higher selfing rate during the process of creation of the base populations, more SSR were simulated in the initial generation (170, instead of 150). We finally had two simulated datasets, termed 'low disequilibria' and 'high disequilibria', with 24 replicates each.

Yangambi oil palm population case study with known pedigree

Simulated populations can be considered as somewhat ideal populations compared to real populations where, for example, selection could have been applied. Therefore, FT* was also tested with a real oil palm dataset.

The Yangambi population originated from the Democratic Republic of Congo with the plantation in the 1920s of palms coming from open pollinations of a few founders

considered as unrelated (Corley and Tinker 2003). The subsequent pedigree that started at that point is well known, with a molecular assessment by SSR markers (Cochard 2008) for families that belong to the CIRAD/PalmElit breeding program (http://www.palmelit.com). To validate FT*, we used data from 16 oil palms of these families. Their genealogy covered four generations, back to founder individuals (Fig. 1). This genealogy included selfings from the second generation onwards and one individual was used as male and female in different matings. Moreover, the contribution of individuals to the following generation varied under the effect of artificial selection. The individuals were genotyped with 166 SSR markers.

In order to investigate the effect of the number of SSR on the genealogical coancestries estimated by FT*, we tested seven numbers of SSR (from 6 to 166). For each number of SSR, we conducted eight repetitions, by randomly choosing subsets of SSR.

Measures of accuracy

In order to measure the accuracy of the pedigree reconstruction when validating the approach with either simulated or real data, the Pearson correlation and the root mean square error (RMSE) were calculated between the true and estimated genealogical coancestries, as described by Fernández and Toro (2006).

Application to the Deli oil palm population with scarce pedigree data

The Deli population originated from four oil palms planted in 1848 in Indonesia and the first selfings were done in the 1920s (Corley and Tinker 2003). It is now a key oil palm breeding population, as it has a high agronomic value and because its complementation with other populations (in particular La Mé) leads to heterosis in total bunch production (Cochard et al. 2009). Some knowledge on the history of this population is available but detailed genealogical information only exists for the recent past. FT* was applied to 104 oil palms of the last generation of the Deli population used in the CIRAD/PalmElit breeding program. Their pedigree was mostly unknown and the depth of the known part of their pedigree varied among families for up to a maximum of four generations (Fig. S1, available as Supplementary Data). The unknown part of the pedigree of those 104 individuals was reconstructed from their molecular data, back to the four founders. The 104 individuals were genotyped with 160 SSR. In accordance with breeders' knowledge on the history of the population, FT* was

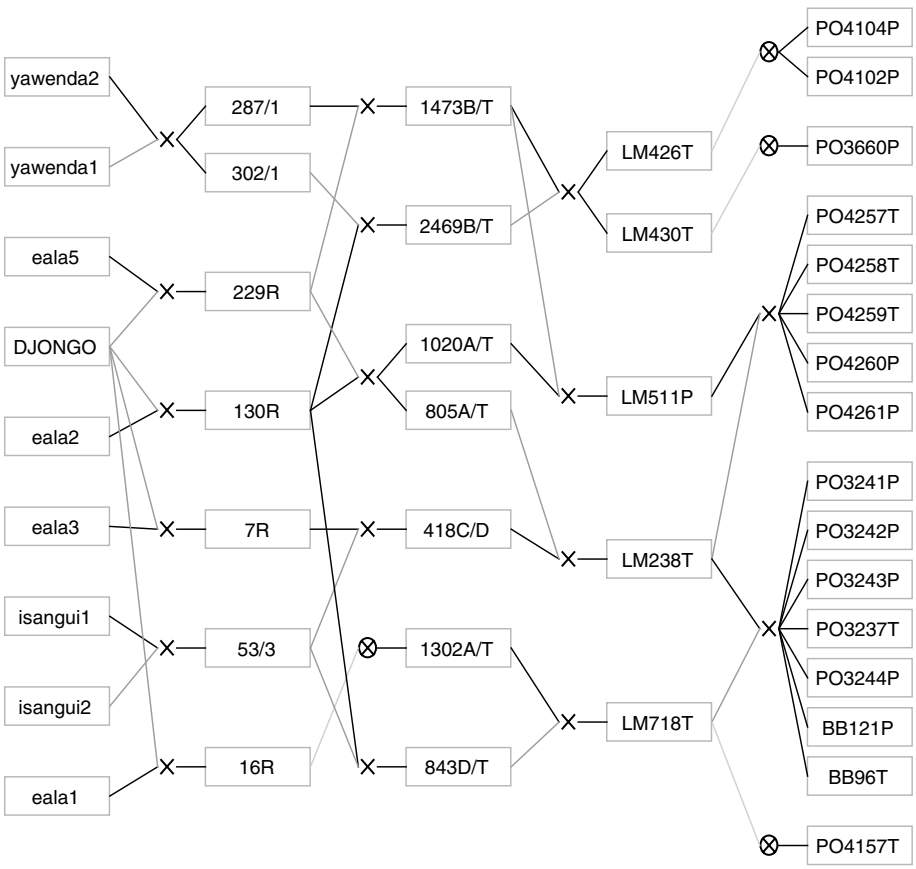


Fig. 1 Pedigree of the Yangambi population of oil palm (with first generation on the *left* and fourth generation on the *right*). The nine founders were considered unrelated. Pedimap software (Voorrips 2007) was used to produce this figure

run alternatively with seven and nine generations elapsed between founders and studied individuals. In addition, the maximum number of individuals per past generation was set in order not to be limiting for the corresponding number of offspring in the known part of the pedigree. Here, all the genotyped individuals did not exactly belong to the same generation compared to the founders, as in the recent past some families had been submitted to more breeding generations. However, this was not a problem for FT* as it concerned the known part of the pedigree, which was arranged so that the most remote known ancestors of each family were in the same generation compared to founders, i.e., in the pedigree some recent generations were skipped for families submitted to fewer breeding generations (see Fig. S1). We tested the statistical power of the methodology according to the number of SSR using 11 levels of SSR numbers (from 8 to 160). For each number of SSR, 16 repetitions were made, by randomly choosing subsets of SSR. As the true genealogical coancestries were unknown, we only measured the fit between the estimated genealogical coancestries and molecular coancestries with RMSE and Pearson correlation.

When estimating genealogical coancestries, FT* reconstructed a pedigree that was compatible with the observed diversity and the molecular relationship between genotyped individuals. Therefore, the estimated genealogy could be used to calculate historical effective population sizes (*Ne*), which is the size of an idealized Wright–Fisher population that would give rise to the same extent of random genetic drift as the actual population (Caballero 1994). Here we used the reconstructed pedigree to estimate *Ne* with the pedigree-based approach developed by Gutiérrez et al. (2008, 2009) and Cervantes et al. (2011). This approach estimates the realized inbreeding and coancestry *Ne* from the individual increase in inbreeding, which is computed for each individual starting from its most remote ancestors. This *Ne* accounts for differences in the depth of the pedigree between lineages and also for all departures between the real and ideal conditions due, for instance, to selection. A modified version of ENDOG software 4.8 (Gutiérrez and Goyache 2005) which accounts for self-fertilization, was used to calculate realized *Ne* in the Deli breeding population from the reconstructed pedigree. As a control, ENDOG was also applied to the real pedigree of the Yangambi population as well as to its reconstruction (made with all SSR). In order to compare the results with a method independent of the pedigree data, we also used the approach of Hill (1981) and Waples (2006), which calculates inbreeding *Ne* in the parental generation from linkage disequilibrium between unlinked markers. LDNE software version 1.31 (Waples and Do 2008) was used for this task. Calculations were performed with three sets of 16 SSR chosen on different linkage groups, according to the reference map of Billotte et al. (2005).

Results

Testing the effect of selfing rate and marker numbers with simulated data

As can be seen in Fig. 2a, the method gave the best results for selfing rates below 0.6 and at least 30 SSR. In these conditions, the root mean square error (RMSE) between the true and estimated genealogical coancestry was small (<0.07). When the selfing rate passed 0.6, the RMSE also increased and finally reached 0.16. The RMSE was the same with 30 or 90 SSR, but was significantly larger with 10 SSR. The evolution of the linear regression line between the estimated and true genealogical coancestries according to the selfing rate helps in analyzing the RMSE profile (Fig. 2b). The decreasing regression slope with increasing selfing rate indicated that the genealogical coancestries were misestimated, with bias increasing with the selfing rate, especially for closely related individuals in the true pedigree. When the selfing rate reached one, the bias became strong as the slope was only 0.55, which coincided with the high observed RMSE.

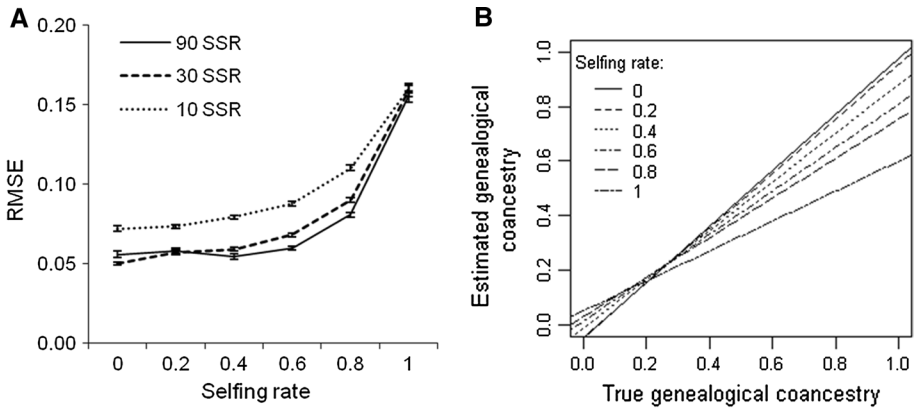


Fig. 2 Effect of the selfing rate in the true pedigree on **a** the root mean square error (RMSE) between estimated and true genealogical coancestries, according to the number of SSR markers (10, 30 and 90), and **b** on the regression between estimated and true genealogical coancestries using 90 SSR. In **a**, *bars* are SEM ($n = 50$). In **b**, each line is the average regression of estimated coancestry on true genealogical coancestry over 50 replicates

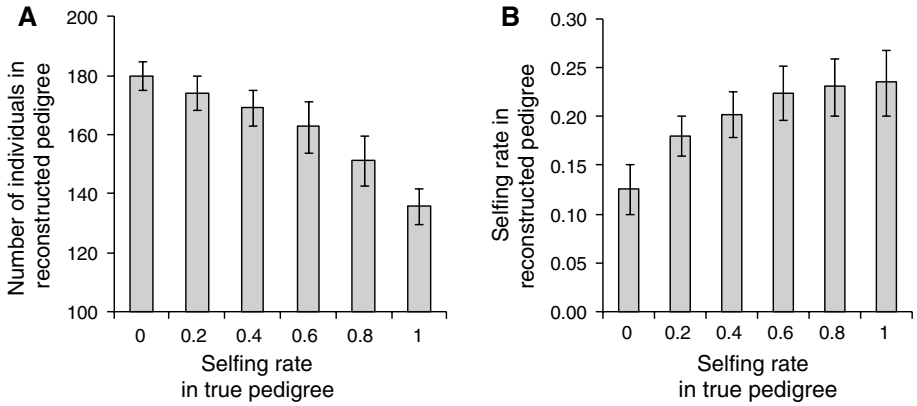


Fig. 3 Effect of the selfing rate in the true pedigree on **a** the number of individuals (true number of individuals = 120) and **b** the selfing rate in the reconstructed pedigree. 90 SSR were used. *Bars* are SD ($n = 50$)

The above performance was a consequence of the selfing rate and also the interaction of this parameter with the number of virtual individuals in reconstructed pedigrees. The method always overestimated the number of individuals in the reconstructed pedigree, from 13 to 50 % (Fig. 3a) and, therefore, this led to underestimated genealogical coancestries. However, when the true selfing rate was under 0.2, this underestimation of coancestries was partly compensated by an overestimation of the selfing rate (Fig. 3b), leading to a small RMSE. When the true selfing rate increased above 0.2, FT* underestimated the selfing rate, as it was unable to generate pedigrees with a selfing rate higher than 0.30, leading to a greater coancestry underestimation.

Surprisingly, the Pearson correlation between the real and estimated genealogical coancestries gave apparently contradictory results, as it increased in parallel with the selfing rate (Fig. S2, available as Supplementary Data). However, the Pearson correlation was actually not relevant to evaluate the effect of an increase in the self-fertilization level. Indeed, higher selfing rates led to more homogeneous and differentiated families, so the estimates had high correlations even when the values were biased (high RMSE). This is illustrated in Fig. S3, available as Supplementary Data.

Testing the effect of the percentage of unknown parentages and marker numbers with simulated data

According to the number of markers used and the percentage of parentages removed, the Pearson correlation between the true and estimated coancestries ranged from 0.8 to 0.969 and the RMSE from 0.033 to 0.074 (Fig. 4a, b). The Pearson correlation increased and the RMSE decreased with the number of markers, especially between 10 and 36 SSR and little change in both parameters was observed with a further increase in the number of markers. As expected, the Pearson correlation decreased and the RMSE increased when a larger part of the pedigree

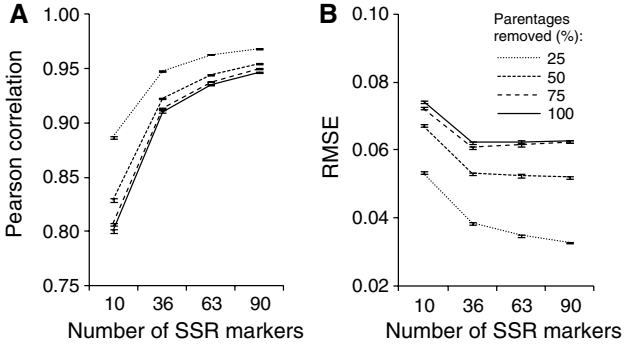


Fig. 4 Effect of the percentage of unknown parentages and number of SSR markers on the **a** Pearson correlation and **b** root mean square error (RMSE) between estimated and true genealogical coancestries. *Bars* are SEM ($n = 704$)

was unknown. When the pedigree was assumed to be completely unknown, the RMSE plateaued at 36 SSR and the Pearson correlation at 90 SSR. Using sets of 30–90 SSR led to a Pearson correlation of over 0.9 and an RMSE under 0.07.

Testing the effect of LD and HW in the base population with simulated data

The mean r^2 measure of LD in the 'NON IDEAL' base populations was 48.4 % higher than in the 'IDEAL' populations in the 'low disequilibria' datasets and 60 % higher in the 'high disequilibria' datasets ($p < 0.001$ for all replicates). The mean p value of the HW test in the 'IDEAL' base populations was 8.8 % higher than in the 'NON IDEAL' populations in the 'low disequilibria' datasets and 41.5 % higher in the 'high disequilibria' datasets ($p < 0.05$ for all replicates) (Table 2). As random genetic drift induced allele fixation at some SSR during the process of creation of the base populations, the final number of SSR that were polymorphic in the base population and used for pedigree reconstruction fell to an average of 92 ± 6 (SD)

and 103 ± 8 in the 'low' and 'high disequilibria' datasets, respectively. In the last generation (genotyped individuals whose pedigree was reconstructed), polymorphism for these SSR was low, with an average of 2.2 ± 0.4 (SD) and 2.1 ± 0.4 in the 'low' and 'high disequilibria' datasets, respectively. Finally, the simulation results showed that the status of the base population regarding the HW and linkage equilibrium had no effect on the genealogical coancestry estimation, as both the RMSE and Pearson correlation were similar with the two base populations, regardless of the strength of the departure from the ideal conditions (Table 2).

Yangambi oil palm case study with known pedigree

The Pearson correlation increased and the RMSE decreased with the number of markers, especially between 6 and 38 SSR, and little change in both parameters was observed with a further increase in the number of markers (Fig. 5a, b). Using 38 SSR led to a Pearson correlation of above 0.9 and an RMSE under 0.08. This result was in agreement with the simulation results.

Selfing occurred in this population at a rate of 0.18, while the rate estimated from the reconstructed pedigree was 0.27 when using all markers. This discrepancy in the selfing rate between the true and reconstructed pedigree was consistent with the simulation results obtained with a true selfing rate close to 0.20. The number of individuals in the pedigree was overestimated by 34.1 %, which was also consistent with the simulation results.

When using FT* with a real dataset, the only summary statistic is the Pearson correlation between the molecular and estimated genealogical coancestries. To check if this statistic could be used as a measure of error of the true pedigree, we compared its evolution with the evolution of RMSE and Pearson correlation between the true and estimated genealogical coancestries according to the number of SSR. The Pearson correlation between the molecular and estimated genealogical coancestries was high even with 6 SSR (0.95 ± 0.01 SD) and reached a plateau at 70 SSR (not shown). Therefore, when applying FT* the effect of the number of markers must be investigated in order to know if enough markers were used to achieve the best possible result that the method can yield for the dataset. This point is crucial, as a small increase in the Pearson correlation between the molecular and estimated genealogical coancestries can be associated with a strong increase in the quality of the pedigree reconstruction. Thus, in the Yangambi

Table 2 Effect of the base population on genealogical coancestry estimation

Base population	r^2 (LD)	p value (HW)	RMSE	Pearson
IDEAL	0.031 ± 0.002	0.626 ± 0.043	0.112 ± 0.019	0.739 ± 0.038
NON IDEAL (low disequilibria)	0.046 ± 0.003	0.576 ± 0.048	0.109 ± 0.018	0.742 ± 0.035
IDEAL	0.035 ± 0.002	0.638 ± 0.024	0.099 ± 0.015	0.777 ± 0.022
NON IDEAL (high disequilibria)	0.056 ± 0.004	0.451 ± 0.061	0.104 ± 0.017	0.754 ± 0.029

Values are mean ± SD ($n = 24$). r^2 is the measure of linkage disequilibrium (LD) obtained by the squared Pearson correlation between markers. p Value (HW) comes from exact tests for Hardy–Weinberg departure. The root mean square error (RMSE) and the Pearson correlation were calculated between estimated and true genealogical coancestries

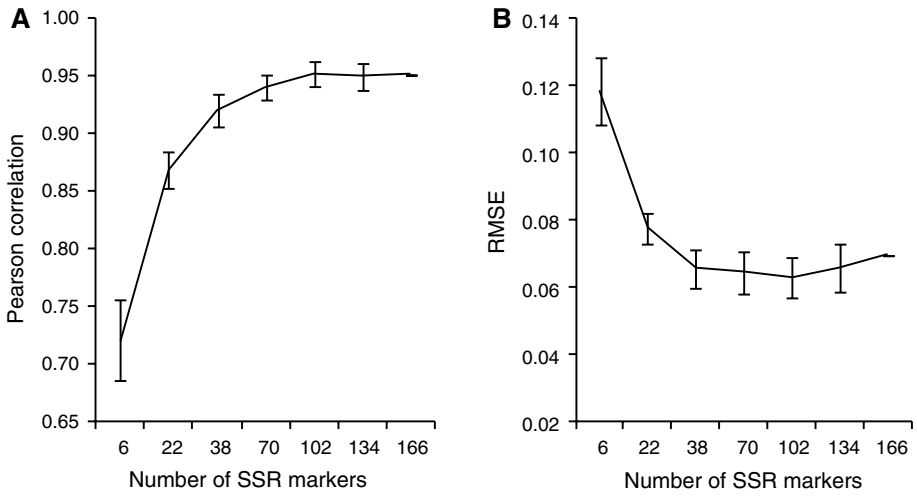


Fig. 5 Effect of the number of SSR markers on the **a** Pearson correlation and **b** root mean square error (RMSE) between estimated and true genealogical coancestries in the Yangambi oil palm breeding population. *Bars* are SEM ($n = 8$)

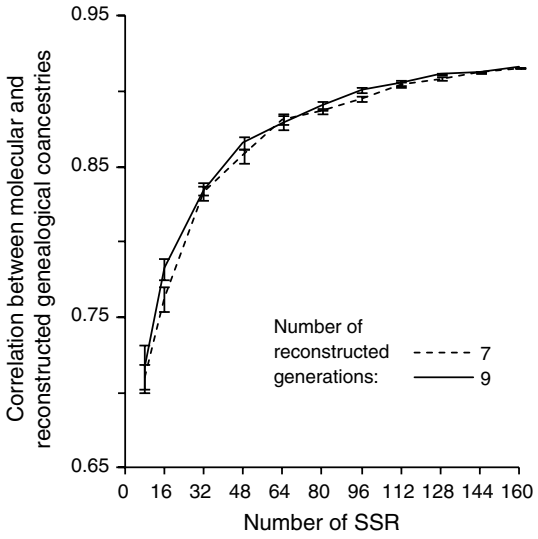


Fig. 6 Effect of the number of SSR markers and the number of generations to reconstruct on the Pearson correlation between molecular and estimated genealogical coancestries in the Deli oil palm breeding population. *Bars* are SEM ($n = 16$)

dataset, while the Pearson correlation between the molecular and estimated genealogical coancestries increased from 0.95 to 0.98, the Pearson correlation between the true and estimated genealogical coancestries increased from 0.72 to 0.95 and the RMSE decreased from 0.12 to 0.07.

Application to the Deli oil palm population with scarce pedigree data

According to the number of markers used and the depth of the reconstructed pedigree, the Pearson correlation between the molecular and estimated genealogical coancestries ranged from 0.710 to 0.916 (Fig. 6). The Pearson correlation increased with the number of SSR according to a diminishing return trend. For all numbers of SSR, the Pearson correlation was similar, with seven and nine reconstructed generations, with slightly higher levels for the latter case but with no significant differences. Fernández and Toro (2006) already found that the correlation between the true and estimated genealogical coancestries increased with the number of virtual generations, even when the true genealogy had less generations than the estimated one. A high Pearson correlation between the estimated genealogical coancestries and molecular coancestries (>0.9) was achieved when using more than 100 markers.

For seven and nine generations, the realized inbreeding *Ne* was 4.3 ± 1.3 (SD) and 5.3 ± 1.3 and the ratios of the coancestry *Ne* to the inbreeding *Ne* were 2.0 and 1.8, respectively. LDNE gave an inbreeding *Ne* of 5.0 ± 1.1. In the Yangambi population, the realized inbreeding *Ne* from the reconstructed pedigree (10.9 ± 4.3) and the true pedigree (7.1 ± 2.4) were not significantly different, indicating that using the reconstructed pedigree to estimate *Ne* was relevant.

Discussion

When the pedigree is unknown or scarcely known, it could be worthwhile to recover or estimate the genealogical relationship between available individuals on the basis of molecular information. Fernández and Toro (2006) method (FT) exhibits good properties compared to other methods reported in the literature. However, FT was originally designed to deal with data from species with separate sexes and, therefore, it was not applicable to many plant species. In the present study, we have adapted the FT method to deal with hermaphroditism and monoecy, with the possibility of selfing. In addition, some improvements were made to the method to take previous knowledge on the population demographic history into account. The new version of the method (FT*) showed good performance on simulated as well as real data of mixed-mating species using a realistic number of markers. The influence of different parameters on the accuracy of the method was also determined.

Accuracy and current limitations of the modified FT method

The FT* method overestimated the number of individuals in the genealogy, while the selfing rate could be either over- or underestimated, depending on the true selfing rate. The best results were obtained for a selfing rate of below 0.6, which shows that this method is suitable for many plant populations. For instance, Jarne and Auld (2006) reported that around 60 % of hermaphroditic plant species had a selfing rate below 0.4. However, when the true selfing rate is likely to be very high or very low, it would be interesting to use this information as a constraint in the pedigree reconstruction. Consequently, further improvements of the method could include a user-specified percentage of selfing for any of the virtual ancestor generations.

With our simulated data, the new method gave similar results to those presented in Fernández and Toro (2006) for unbalanced simulated datasets of small populations. When using enough markers, they also achieved a small RMSE. Moreover, our real Yangambi oil palm data and the real pig data used by Fernández and Toro (2006) yielded similar results. As we could expect, both datasets showed that the number of markers was a limiting factor, since the reliability of the coancestry estimates was markedly reduced when it was too small.

Marker polymorphism appeared to be another key parameter interacting with the number of markers required

to get the best results. In the pig dataset used by Fernández and Toro (2006), one marker per chromosome with 4.2 alleles per marker gave results very close to the best values obtained (using all markers), which was also the case in our first and second sets of simulations using only 10 SSR. In our third group of simulations, although an average number of 92 SSR were used, the coancestry estimation did not achieve the same quality as in the first and second simulations, as indicated by the higher RMSE and lower Pearson correlation. This could be related to the marker polymorphism, which was lower in the third simulation (an average of 2.2 alleles per SSR). Finally, when the number of markers was sufficiently high (from 30 to 100, according to their polymorphism), a high Pearson correlation and low RMSE between true and estimated coancestries were obtained.

The maximum number of individuals in past generations also played a very important role in the quality of the coancestry estimation. An artificially enlarged number of ancestors leads to a larger feasible space of solutions thus making it harder for the underlying optimization algorithm to find the fittest solution. Here, the maximum number of individuals per past generation was at least twofold higher than the true value, which led FT* to substantially overestimate the number of individuals. Clearly, if the maximum number of past individuals could be accurately defined (roughly, less than one-third higher than their true number), the overestimation of the number of individuals and, consequently, the selfing rate bias would have been lower. All of these considerations highlight the importance of including correct information on the past demographic structure of the population to get accurate estimates, as close as possible to the true pedigree.

As FT* makes no assumptions about Hardy–Weinberg and linkage equilibria in the base population, its results were expected to be unaffected by possible departure from these ideal conditions. We confirmed this point here. This is a very important feature of FT, as many methods used to infer relationships from molecular markers, either by explicit pedigree reconstruction or using pairwise estimators, make these assumptions (Fernández and Toro 2006). Moreover, as the number of markers increases rapidly (through SNP panels or Next Generation Sequencing), it will be common to have markers in linkage disequilibrium.

The FT* method is only suitable for diploid species. It had not been a concern in the original version as polyploid animal species are rare. It is however much more common in plant species, as many important crops are polyploid. Clearly, the current version could be extended to polyploid species. The method to calculate molecular coancestries should be modified and changes should be made in the rules to check for molecular incompatibilities within full-sib families, but this would be rather straightforward. From an operational standpoint, the program should account for individuals carrying more than two alleles per loci and adequately simulate the way of transmission of genetic information from one generation to the next.

As stated before, FT* only considers discrete generations of virtual ancestors. Thus, it assumes that the parents of an individual always belong to the previous generation and that mating between two individuals of different generations is not possible. This point can be limiting, for instance for natural populations of perennial species or for breeding populations where clonally propagated individuals have been used repeatedly throughout the pedigree. Notwithstanding, the objective of the method is to estimate a genealogical coancestry matrix between the available individuals at a particular time (e.g., selection candidates in a breeding program), not necessarily to reconstruct the exact real pedigree leading to them. Consequently, the aim is to get a pedigree which is compatible with the observed structure of molecular relationships and that implies the same level of genetic drift. Another limitation of the method is that it assumes that all available individuals with a genotype belong to the last generation, while molecular data can also be available for individuals of past generations even if they are no longer present in the population. To account for both situations, the objective would be somewhat different, i.e., to reconstruct the true pedigree, at least for all genotyped individuals. Modifying the method to allow for the use of molecular data over several generations would require checking the molecular compatibility between relationships apart from full-sib families and would also require using other information such as the date of birth of each individual or the age when they were reproductively mature. This task would be computationally more complex than in the present situation.

Few methods explicitly reconstruct a multigeneration pedigree from molecular data. Almudevar (2003) and Riester et al. (2009) used a simulated annealing algorithm to find the maximum likelihood pedigree. Cowell (2009) developed an exhaustive search algorithm adapted from a Bayesian network learning algorithm. Almudevar (2007) used a computationally intensive fully Bayesian approach to infer a pedigree graph from molecular data. However, all these methods applied to complete samples of individuals, i.e., that all the individuals appearing in the final reconstructed pedigree had molecular data, and possible unsampled parents were assumed unrelated to the others. To our knowledge, only two approaches [FT* and the method of Gasbarra et al. (2007a, b)] apply to genotyped individuals belonging to a single generation. Like FT*, the method of Gasbarra et al. (2007a, b) reconstructs a pedigree from molecular data of contemporaneous individuals and information about the population history (number of generations from founders and approximate size of the population at each generation), assuming nonoverlapping

generations. It differs from FT* as it is based on a Bayesian approach using a Markov chain Monte Carlo algorithm. An interesting feature of this method is that it models both the pedigree and gene flows from the founders down to the genotyped individuals. This ensures molecular compatibility at the pedigree level, not only within full-sib families. However, it requires more data than FT*. Mating parameters controlling the distribution of offspring among males and the degree of monogamy (or estimated from the data), as well as allele frequencies in the base population must be known. A further study is needed to compare the results of those two approaches for the same dataset.

Future uses of FT*

In our study, we focused on breeding populations as we considered that the method would give the best results in these situations. Indeed, their pedigree is generally only partly missing and the known part of the pedigree reduces the possibilities, thus helping the FT* method to approximate the correct relationships. Furthermore, a lot of other information is often available and would further reduce the number of feasible solutions. For example, the structures of some breeding populations contain just full-sib families, or the maximum number of male parents can be established, for instance in polycross designs. However, FT* could also be applied to natural populations. This was already the case with the original FT method, which was applied to several natural animal populations. For example, Zub et al. (2012) estimated heritability in a wild weasel population by applying an animal model using the pedigree reconstructed by FT. Therefore, FT* could be used to study the genetic structure of natural plant populations, evaluate the genetic diversity they harbor and/or design effective management strategies. The FT* method could also be of help in the management of conserved populations, as Fernández et al. (2005) showed that combining molecular markers and pedigree information was useful for increasing the effective population size. However, investigations should be conducted to determine whether combining the molecular coancestry and the genealogical coancestry estimated from markers could give better results than using the molecular coancestry alone, as they could provide redundant information.

Application to the Deli oil palm population with scarce pedigree data

Our use of the reconstructed pedigree of the Deli oil palm breeding population was similar to the approach implemented by Cervantes et al. (2011) in animals, where the reconstructed pedigree of ruminant populations was used to estimate their realized effective size. Here, we found that the reconstructed pedigree was appropriate to estimate *Ne*. However, we constrained FT* with three items of genealogical information: number of founders, known pedigree in recent generations and approximate number of generations between founders and genotyped individuals. Therefore, in this case the reconstructed pedigree should be rather close to the true pedigree, and pedigree-based statistics should be reliable. When FT* is used with no prior information about the history of the population, users should be cautious with statistics calculated with the reconstructed pedigree.

The estimation of *Ne* through the realized inbreeding *Ne* and LD methods refer to different periods of time. The realized inbreeding *Ne* is an average over the time period covered by the pedigree, while LDNE gives the *Ne* of the parental generation. The bottleneck event at the founding of the Deli population created LD, which declined thereafter as the population expanded. Therefore, we could expect the realized *Ne* to be lower than the LDNE value. However, Waples (2005) showed that *Ne* based on LD can be underestimated for several generations under the effects of recent bottlenecks and population size increases. Therefore, the LDNE value could have actually been affected by the recent history of the population, and also reflect the *Ne* before the parental generation. In any case, the results obtained with these two methods in our data appeared to be consistent with each other.

This is the first report of *Ne* in oil palm. Meuwissen (2009) estimated that the critical *Ne* for a population was around 50–100, on order to avoid long-term inbreeding problems. Our results highlighted that efforts should be made to limit inbreeding in the oil palm breeding populations studied and to diversify their genetic base. For example, this could be achieved for the Deli population by crossing it with some African populations, as suggested by Cochard et al. (2009). As the coancestry *Ne* and the inbreeding *Ne* are measures of the same drift process, they should reach an identical asymptotic value in an idealized population without permanent sublining. Therefore, their difference gives an estimate of preferential matings. In the Deli population, the ratio of the coancestry *Ne* to the inbreeding *Ne* indicated a high degree of subdivision due to nonrandom mating. This could likely be explained by the selection applied to this population, which first underwent mass selection and reciprocal recurrent selection afterwards. This was also reflected in the high variability in the average relatedness of the four founders, which is a measure of their genetic contribution to the population (Gutiérrez and Goyache 2005). For instance with seven reconstructed generations, the average relatedness given by ENDOG ranged from 8.3 to 44.2 %.

Conclusion

The FT* method gave reliable coancestry estimates for mixed-mating species, especially when the selfing rate was lower than 0.6, using a realistic number of markers. We confirmed that the existence of linkage disequilibrium and departure from the Hardy–Weinberg equilibrium in the base population did not affect the method. In a case study, this approach gave valuable information about the Deli oil palm population. This highlighted the potential benefits that plant breeders could obtain by looking for new tools in the animal breeding sector and adapting them to their own circumstances. The method was implemented in the software MOLCOANC 3.0, where the user can choose between separate sexes and mixed-mating. The program is available from the web page http://dl.dropbox.com/u/5714008/Fernandez.htm.

Acknowledgments We would like to thank Virginie Pomiès (CIRAD, Montpellier) for her technical assistance in genotyping, Dr Juan Pablo Gutiérrez (Universidad Complutense de Madrid) for his help with ENDOG software and the anonymous reviewers for their helpful comments. This research was partly funded by a grant from PalmElit SAS.

Conflict of interest The authors declare no conflict of interest.

References

Adams W, Neale D, Loopstra C (1988) Verifying controlled crosses in conifer tree-improvement programs. Silvae genetica 37:147–152

Almudevar A (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theor Popul Biol 63:63–75

Almudevar A (2007) A graphical approach to relatedness inference. Theor Popul Biol 71:213–229

Atkin FC, Dieters MJ, Stringer JK (2009) Impact of depth of pedigree and inclusion of historical data on the estimation of additive variance and breeding values in a sugarcane breeding program. Theor Appl Genet 119:555–565

Billotte N, Marseillac N, Risterucci AM et al (2005) Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). Theor Appl Genet 110:754–765

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evolut Pers Ed 18:503–511

Butler K, Field C, Herbinger CM, Smith BR (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. Mol Ecol 13:1589–1600

Caballero A (1994) Developments in the prediction of effective population size. Heredity 73:657–679

Cervantes I, Goyache F, Molina A, Valera M, Gutiérrez JP (2011) Estimation of effective population size from the rate of coancestry in pedigreed populations. J Anim Breed Genet 128:56–63

Cochard B (2008) Etude de la diversité génétique et du déséquilibre de liaison au sein de populations améliorées de palmier à huile (*Elaeis guineensis* Jacq.). Montpellier SupAgro, Montpellier, p 175

Cochard B, Adon B, Rekima S et al (2009) Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. Tree Genetics Genomes 5:493–504

Corley RHV (2005) Illegitimacy in oil palm breeding—a review. J Oil Palm Res 17:64–69

Corley RHV, Tinker PB (2003) Selection and breeding. In: Blackwell Science Ltd (ed) The oil palm, 4th edn. Blackwell Publishing, Oxford, pp 133–199

Cowell RG (2009) Efficient maximum likelihood pedigree reconstruction. Theor Popul Biol 76:285–291

Doerksen TK, Herbinger CM (2010) Impact of reconstructed pedigrees on progeny-test breeding values in red spruce. Tree Genetics Genomes 6:591–600

Eding H, Meuwissen THE (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159

Emigh T (1980) Comparison of tests for Hardy–Weinberg equilibrium. Biometrics 36:627–642

Emik LO, Terrill CE (1949) Systematic procedures for calculating inbreeding coefficients. J Hered 40:51–55

Ericsson T (1999) The effect of pedigree error by misidentification of individual trees on genetic evaluation of a full-sib experiment. Silvae genetica 48:239–242

Fernández J, Toro MA (2006) A new method to estimate relatedness from molecular markers. Mol Ecol 15:1657–1667

Fernández J, Villanueva B, Pong-Wong R, Toro MA (2005) Efficiency of the use of pedigree and molecular marker information in conservation programs. Genetics 170:1313–1321

Garbe JR, Da Y (2008) Pedigraph user manual Version 2.4. Department of Animal Science, University of Minnesota

Gasbarra D, Pirinen M, Sillanpaa M, Arjas E (2007a) Estimating genealogies from linked marker data: a Bayesian approach. BMC Bioinforma 8:411

Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007b) Estimating genealogies from unlinked marker data: a Bayesian approach. Theor Popul Biol 72:305–322

Guo SW, Thompson EA (1992) Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372

Gutiérrez JP, Goyache F (2005) A note on ENDOG: a computer program for analysing pedigree information. J Anim Breed Genet 122:172–176

Gutiérrez JP, Cervantes I, Molina A, Valera M, Goyache F (2008) Individual increase in inbreeding allows estimating effective sizes from pedigrees. Genet Sel Evol 40:359–378

Gutiérrez JP, Cervantes I, Goyache F (2009) Improving the estimation of realized effective population sizes in farm animals. J Anim Breed Genet 126:327–332

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. Genet Res 38:209–216

Jarne P, Auld JR (2006) Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. Evolution 60:1816–1824

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220:671–680

Kumar S, Gerber S, Richardson TE, Gea L (2007) Testing for unequal paternal contributions using nuclear and chloroplast SSR markers in polycross families of radiata pine. Tree Genetics Genomes 3:207–214

McIntyre CL, Jackson PA (2001) Low level of selfing found in a sample of crosses in Australian sugarcane breeding programs. Euphytica 117:245–249

Meuwissen T (2009) Genetic management of small populations: a review. Acta Agriculturae Scandinavica Section A Animal Sci 59:71–79

Morrissey MB, Wilson AJ (2010) pedantics: an r package for pedigree-based genetic simulation and pedigree manipulation, characterization and viewing. Mol Ecol Resour 10:711–719

Pemberton JM (2008) Wild pedigrees: the way forward. Proc Royal Soc B Biol Sci 275:613–621

Riester M, Stadler PF, Klemm K (2009) FRANz: reconstruction of wild multi-generation pedigrees. Bioinformatics 25:2134–2139

Voorrips RE (2007) Pedimap: software for visualization of genetic and phenotypic data in pedigrees. Plant Research International, Wageningen

Waples RS (2005) Genetic estimates of contemporary effective population size: to what time periods do the estimates apply? Mol Ecol 14:3335–3352

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv Genet 7:167–184

Waples RS, Do CHI (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. Mol Ecol Resour 8:753–756

Zhao J (2007) gap: genetic analysis package. J Stat Softw 23:1–18

Zub K, Piertney S, Szafranska PA, Konarzewski M (2012) Environmental and genetic influences on body mass and resting metabolic rates (RMR) in a natural population of weasel *Mustela nivalis*. Mol Ecol 21:1283–1293

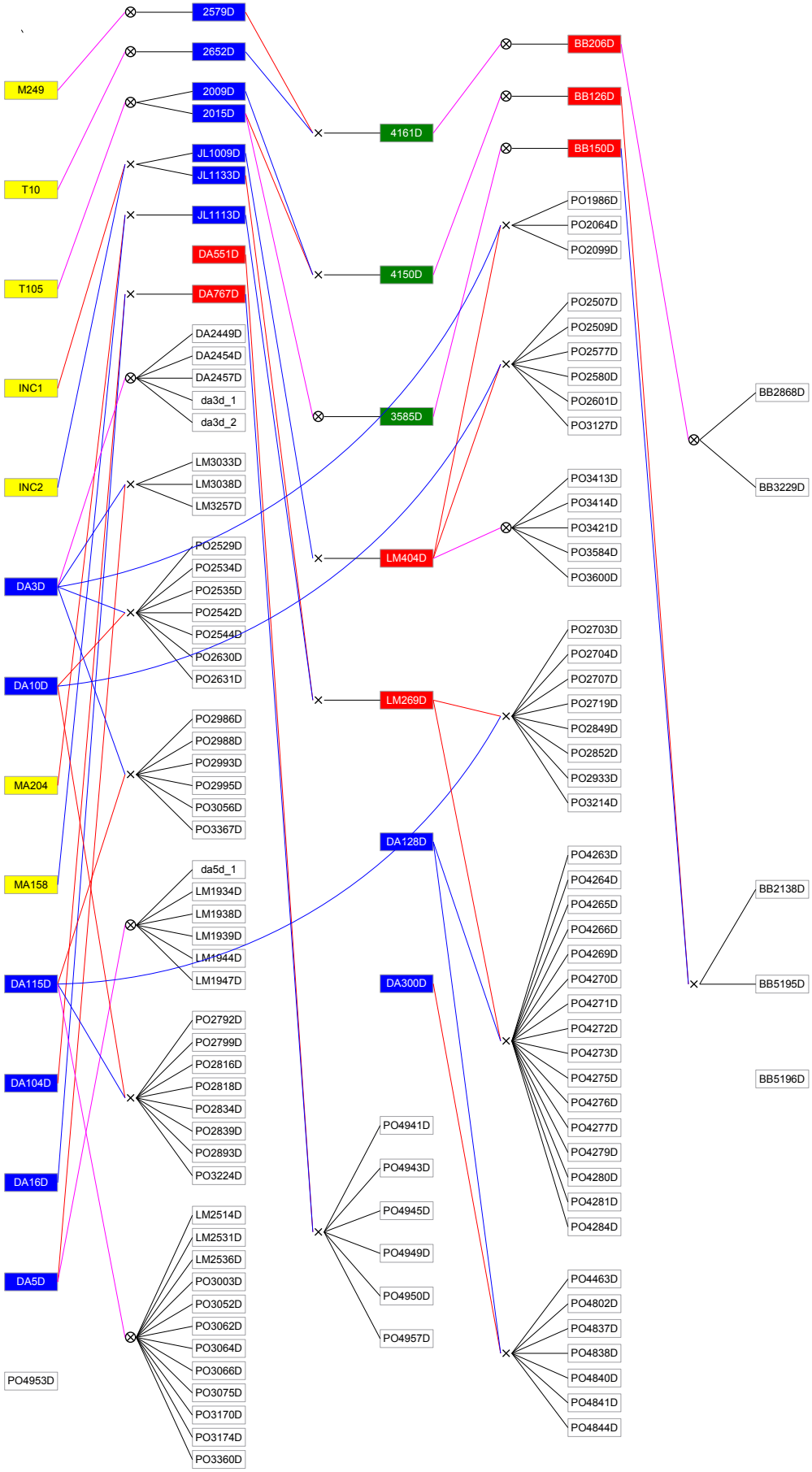


Fig. S1 Known part of the pedigree of the Deli oil palm breeding population used by PalmElit. The pedigree of the genotyped individuals was known for up to four generations, depending on families. The fill color in the blocks indicates the generation, with yellow, blue, green, red and white corresponding to generations 1, 2, 3, 4 and 5, respectively. Molecular data were available for individuals in generation 5. Line colors indicate the type of relationship: red line to female parent, blue line to male parent and pink line to single parent (selfing). Pedimap software (Voorrips 2007) was used to produce this figure

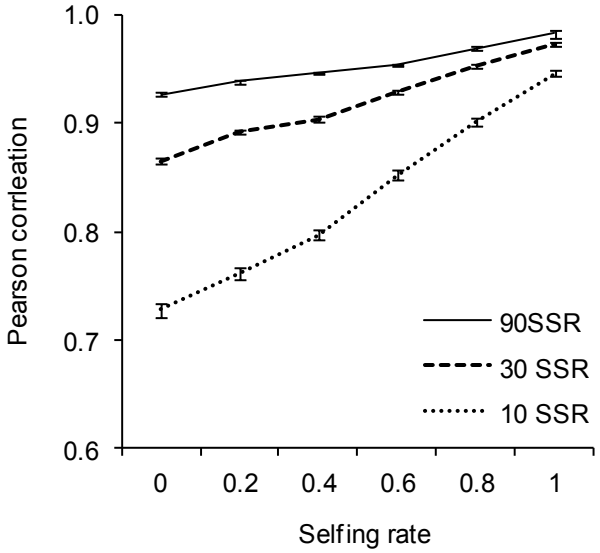


Fig. S2 Effect of the selfing rate in the pedigree on the Pearson correlation between true and reconstructed genealogical coancestries using 10, 30 or 90 SSR. Bars are SEM (n = 50)

Theoretical and Applied Genetics

Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population

David Cros (✉), Leopoldo Sánchez, Benoit Cochard, Patrick Sampers, Marie Denis, Jean-Marc Bouvet, Jesús Fernández.

D. Cros

Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit (AGAP), CIRAD, International campus of Baillarguet TA A-108/C, 34398 Montpellier Cedex 5, France

e-mail: david.cros@cirad.fr

Tel.: +33-467615800

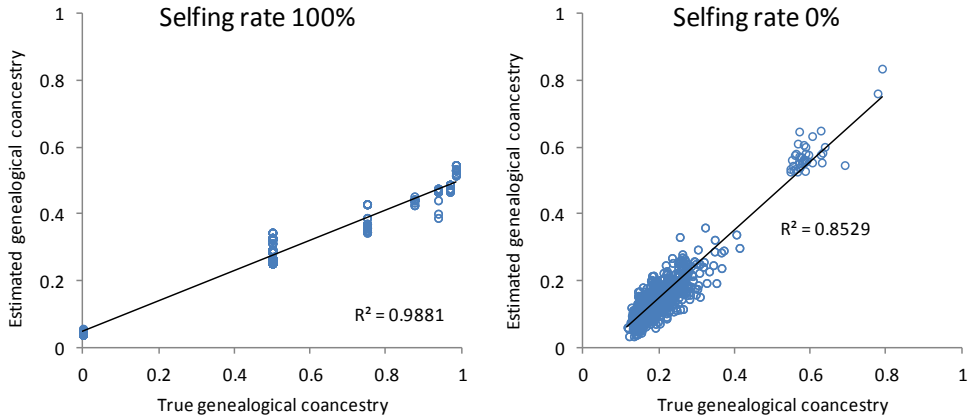


Fig. S3 Correlation between estimated and true genealogical coancestries for two simulated pedigrees of contrasted selfing rates. Each dot is the value of the coancestry between two individuals of the last generation

Theoretical and Applied Genetics

Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population

David Cros (✉), Leopoldo Sánchez, Benoit Cochard, Patrick Sampers, Marie Denis, Jean-Marc Bouvet, Jesús Fernández.

D. Cros

Genetic Improvement and Adaptation of Mediterranean and Tropical Plants Research Unit (AGAP), CIRAD, International campus of Baillarguet TA A-108/C, 34398 Montpellier Cedex 5, France

e-mail: david.cros@cirad.fr

Tel.: +33-467615800

FIGURES ET TABLEAUX

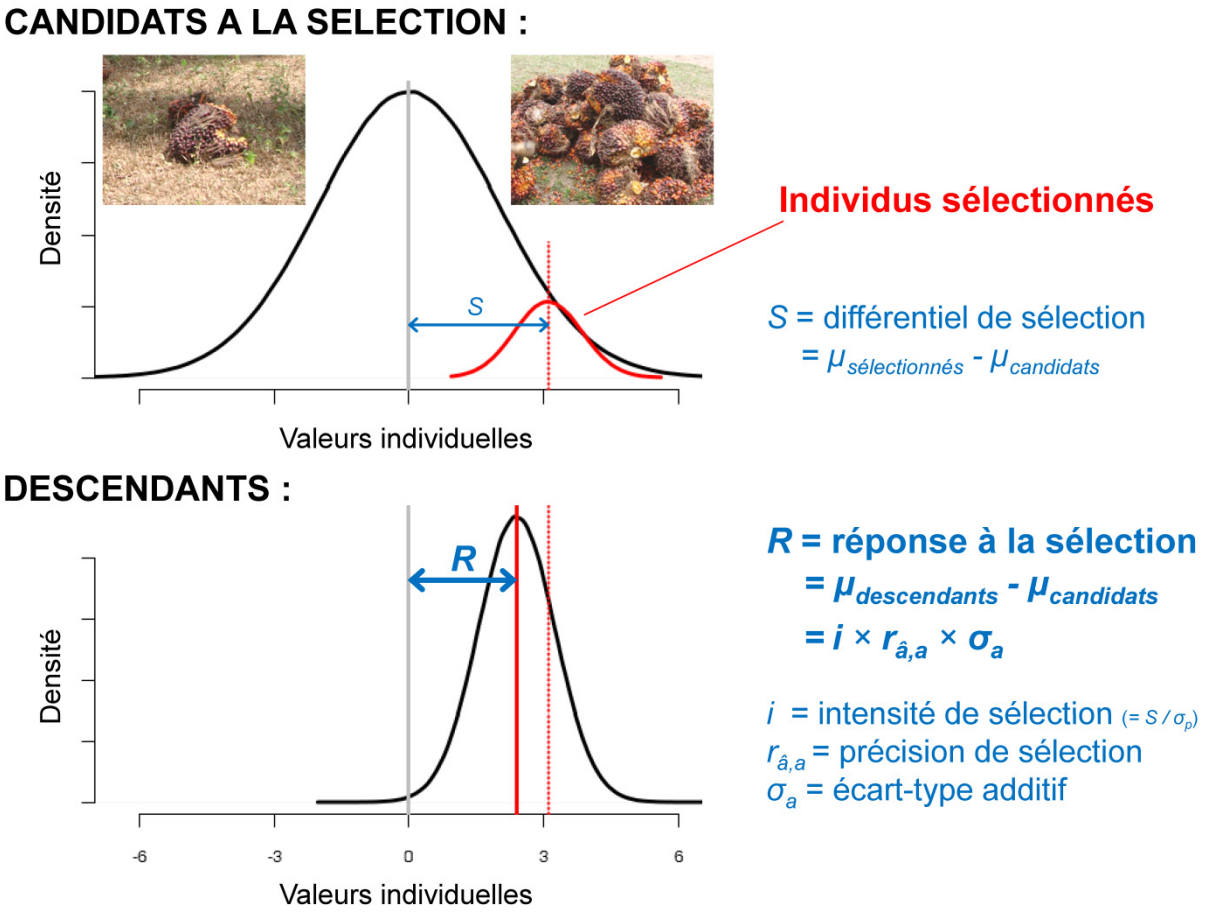


Figure 1 La réponse à la sélection et les paramètres dont elle dépend
La valeur phénotypique des individus sélectionnés (par exemple le rendement en régimes) suit une distribution normale, illustrant une sélection multicaractères (μ = moyenne phénotypique, σ_P = écart-type phénotypique)

Figure 1. Important milestones of selective breeding, which have significantly improved agricultural productivity.

Figure 2 Place de la sélection génomique dans l'histoire de l'amélioration génétique (Jonas et de Koning, 2013)

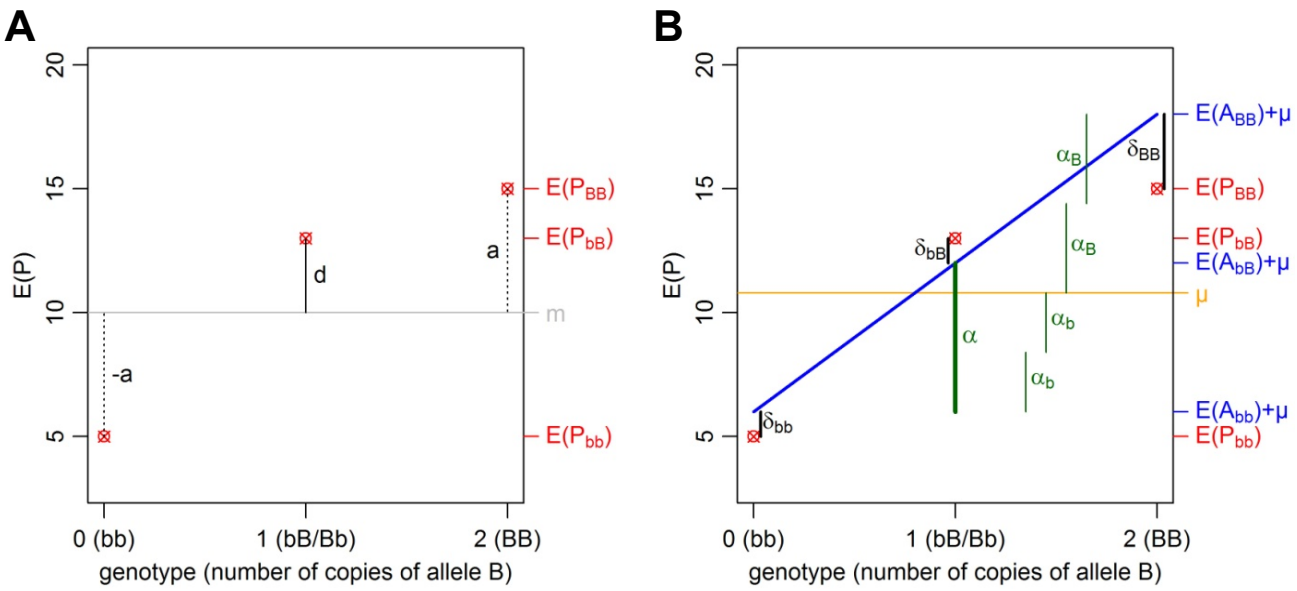


Figure 3 Décomposition de l'espérance phénotypique $P_{..}$ (A) en effets génotypiques additifs *a*, effets génotypiques de dominance *d* et phénotype intermédiaire *m* et (B) en effets moyens des gènes (α_b, α_B), résidus de dominance (δ_{bb}, δ_{bB}, δ_{BB}) et moyenne phénotypique μ dans le cas d'un gène à deux allèles B (favorable) et b (défavorable)

$\alpha_b < 0,\ \alpha_B > 0,\ \alpha = \alpha_B - \alpha_b,\ E(A_{bb}) = \mu + \alpha_b + \alpha_b,\ E(A_{bB}) = \mu + \alpha_b + \alpha_B,\ E(A_{BB}) = \mu + \alpha_B + \alpha_B,$
$E(A) + E(D) = E(G),\ E(G..) + \mu = E(P),\ \delta_{BB} < 0,\ \delta_{bb} < 0,\ \delta_{bB} > 0$

Tableau 1 Coefficient de parenté f_{ij} et de fraternité φ_{ij} et covariance génétique entre deux individus, en l'absence de consanguinité

Individu-…	f_{ij}	φ_{ij}	Covariance génétique
…-individu non apparenté	0	0	0
…-lui-même	0.5	1	$\sigma^2_a + \sigma^2_d$
…-clone	0.5	1	$\sigma^2_a + \sigma^2_d$
…-parent, …-descendant	0.25	0	$0.5\sigma^2_a$
…-plein-frère	0.25	0.25	$0.5\sigma^2_a + 0.25\sigma^2_d$
…-grand-parent, …-petit-enfant	0.125	0	$0.25\sigma^2_a$
…-demi-frère	0.125	0	$0.25\sigma^2_a$
…-premier-cousin	0.0625	0	$0.125\sigma^2_a$

Tableau 2 Moyenne et écart-type (ET) du coefficient d'apparentement ($2f_{ij}$) entre pleins-frères et entre demi-frères en fonction du nombre de loci (d'après VanRaden, 2007)

	Coefficient d'apparentement	
Nombre de loci indépendants	entre pleins-frères (ET)	entre demi-frères (ET)
1	50 (35.4)	25 (17.7)
5	50 (15.8)	25 (7.9)
10	50 (11.2)	25 (5.6)
50	50 (5.0)	25 (2.5)
100	50 (3.5)	25 (1.8)
Infini	50 (0.0)	25 (0.0)

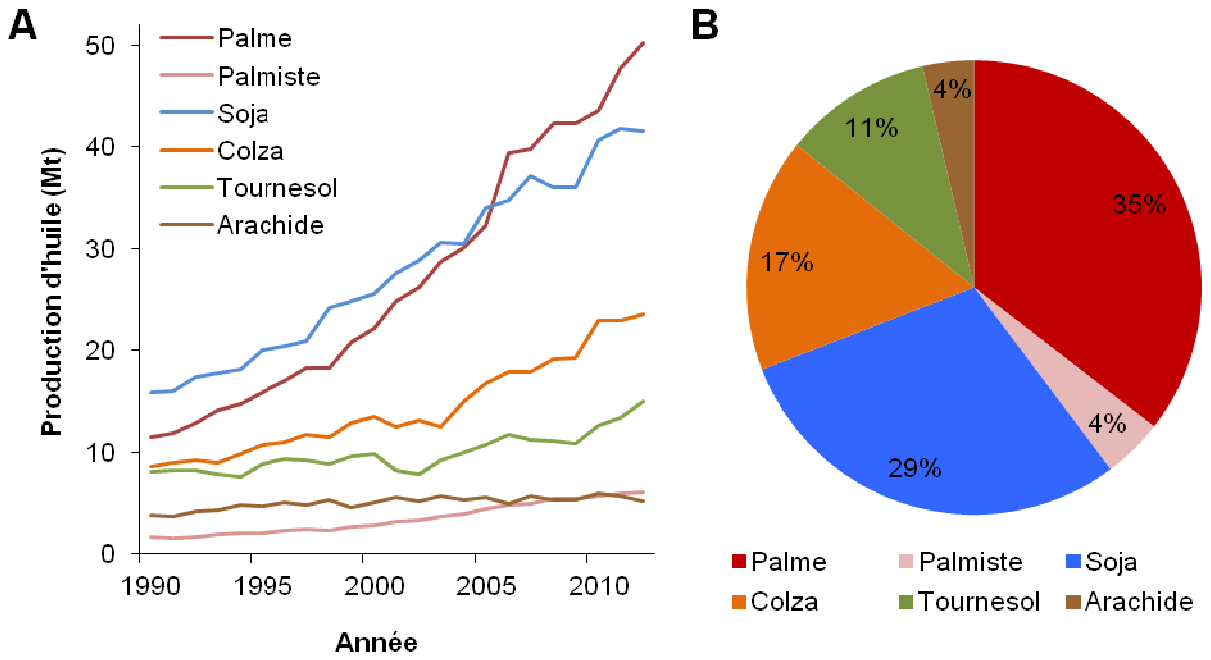


Figure 4 Les principales huiles végétales : (A) Evolution de la production entre 1990 et 2012, (B) Importance relative des différentes huiles dans la production de 2012
(source : FAO, http://faostat.fao.org/, le 16 mai 2014)

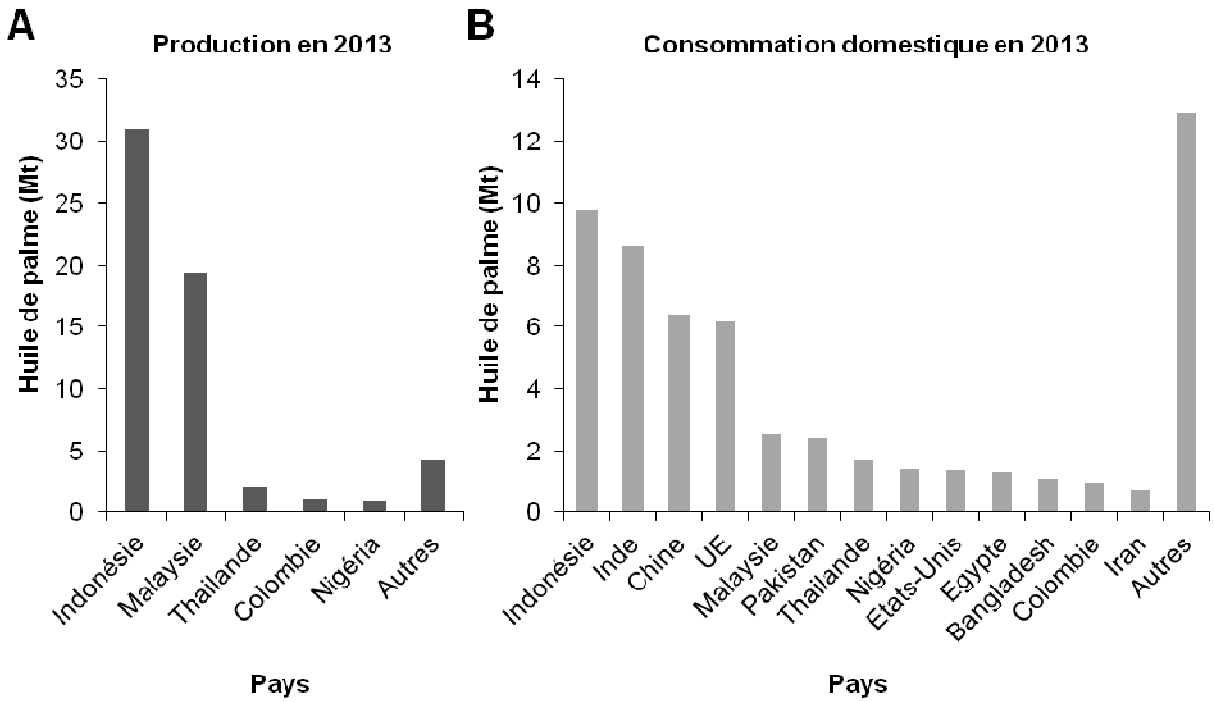


Figure 5 Répartition (A) de la production et (B) de la consommation d'huile de palme entre pays en 2013
(source : USDA, http://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf, le 16 mai 2014)

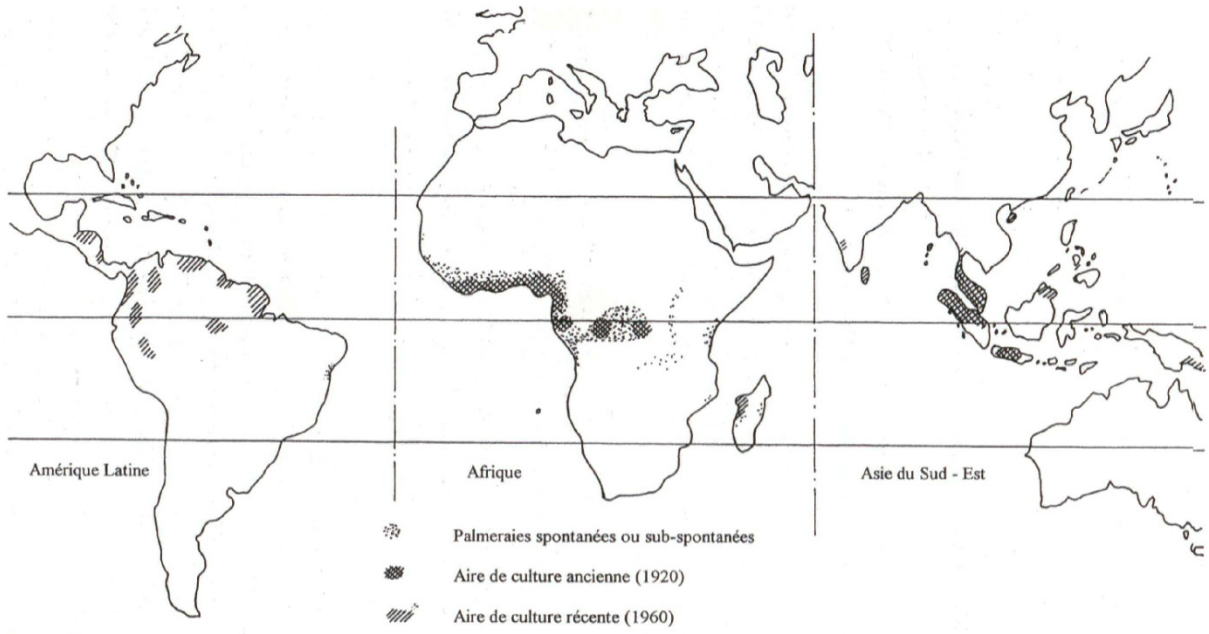


Figure 6 Aire de répartition du palmier à huile (Jacquemard, 1995)

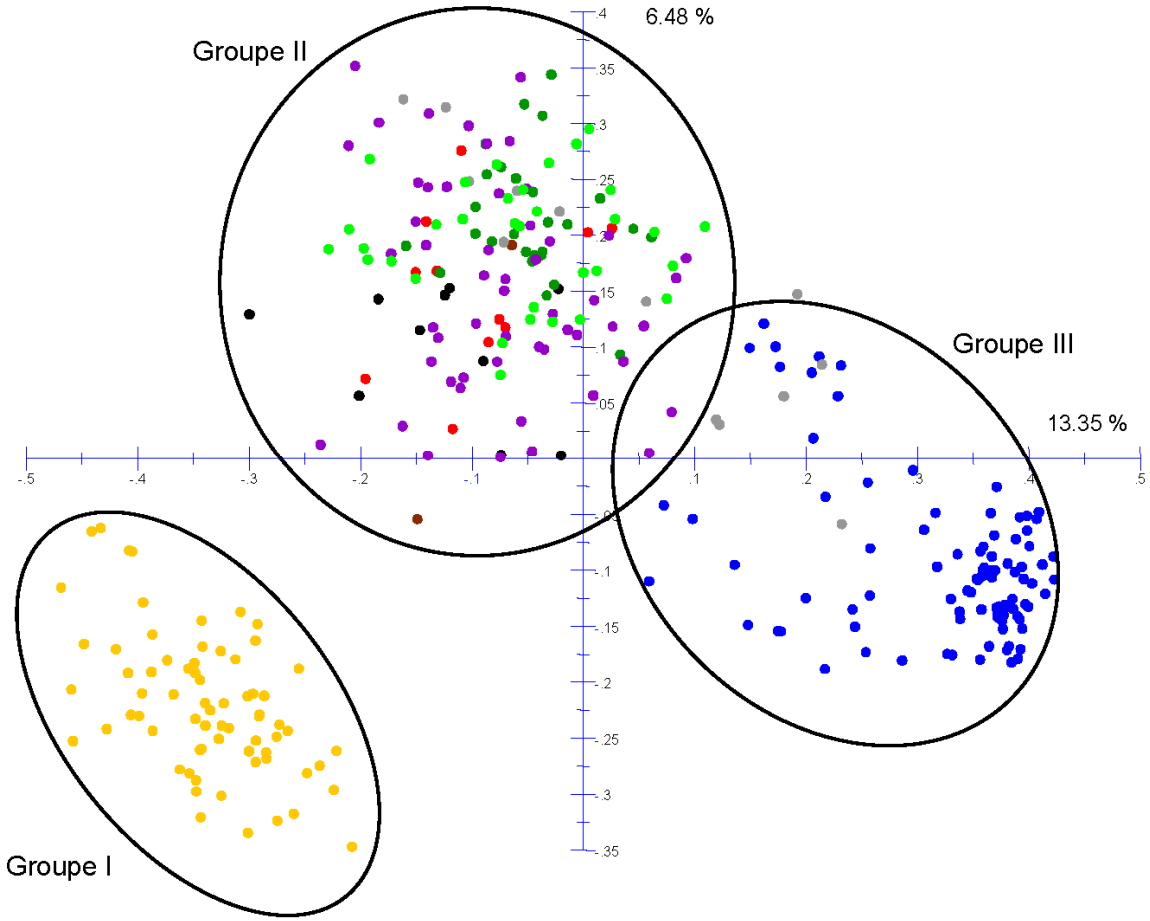


Figure 7 Analyse de diversité génétique au sein d'*E. guineensis* (Cochard, 2008)
Cette étude a été réalisée par analyse en coordonnées principales sur 318 individus génotypés avec 14 microsatellites. Le groupe I rassemble les origines de Côte d'Ivoire. Le groupe II est constitué des origines d'Afrique Centrale plus le Nigeria et Le Bénin. Le Groupe III est composé des origines Deli.

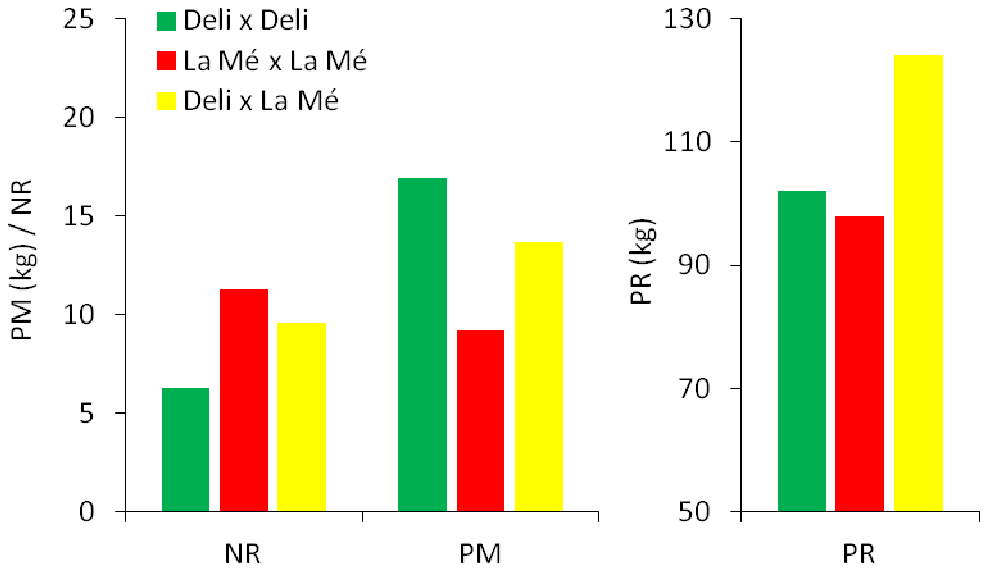


Figure 8 Production totale de régimes (PR) et ses composantes (nombre de régimes NR et poids des régimes PM) à l'âge adulte chez les dura des croisements intra- et inter-populations observés dans « l'Expérience Internationale », d'après les résultats donnés par Gascon et al. (1966)

Le schéma de sélection…

Un cycle de sélection dure au moins 12 ans et comporte 3 étapes :

① Les populations A et B de départ sont soumises à une **sélection sur les caractères les plus héritables** (les moins influencés par l'environnement), comme le pourcentage de pulpe sur fruits.

② Les individus A et B sélectionnés sont croisés entre eux et évalués dans des essais génétiques. La **sélection finale** est faite **sur les aptitudes à la combinaison**.

③ Les individus sélectionnés sont **autofécondés et recombinés** au sein de chaque groupe pour former 2 populations A et B améliorées qui serviront de point de départ au cycle suivant et à produire des **semences commerciales** (④).

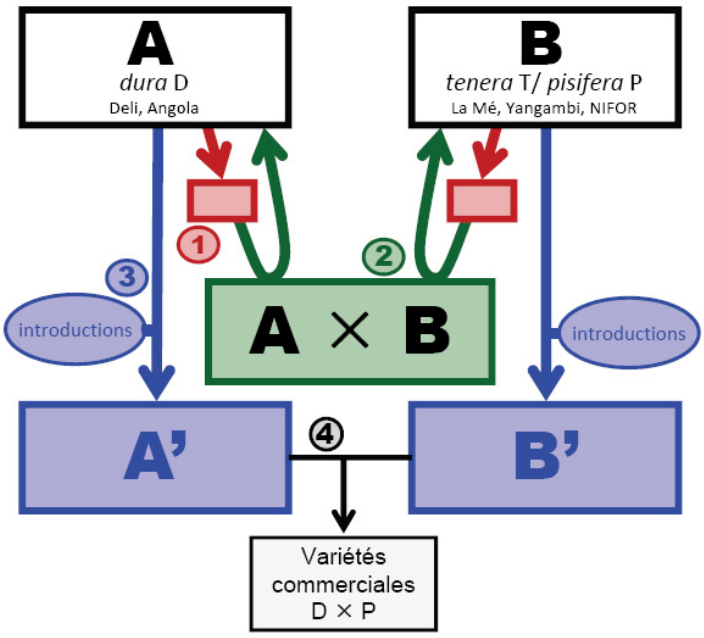


Figure 9 Schéma de la sélection récurrente réciproque appliquée au palmier à huile depuis 1957

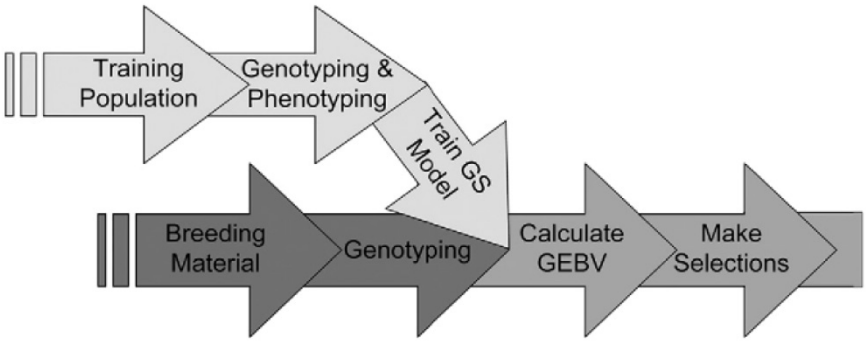


Figure 1. Diagram of genomic selection (GS) processes starting from the training population and selection candidates continuing through to genomic estimated breeding value (GEBV)–based selection. Note that while we show here a single occurrence of model training, training can be performed iteratively as new phenotype and marker data accumulate.

Figure 10 Schéma de principe de la sélection génomique (Heffner et al., 2009)

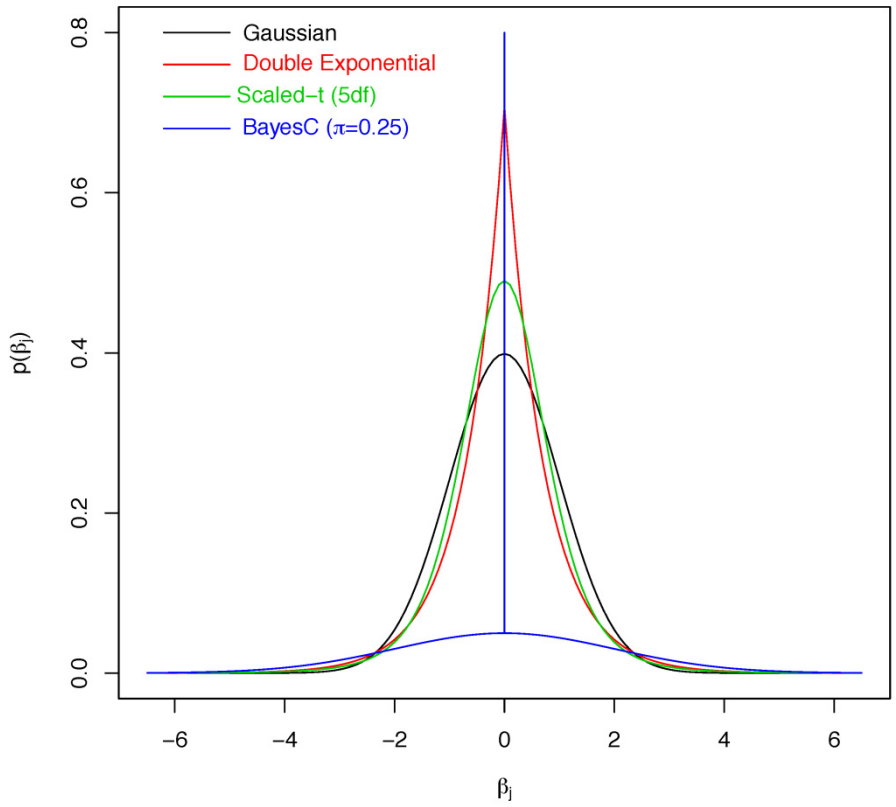


Figure 11 Exemple de distributions a priori des effets aux marqueurs (β_j) pour différentes méthodes bayésiennes de sélection génomique (Pérez et de los Campos, 2013)
Gaussian, en noir : BRR, *Double Exponential*, en rouge : BLR, *Scaled-t,* en vert : BayesA et en bleu : BayesCπ.

Tableau 3 Caractéristiques et propriétés de trois méthodes statistiques de sélection génomiques : RR-BLUP, BayesA et BayesB (Heffner et al., 2009)

Table 1. General characteristics and trends of performance for traditional best linear unbiased predictor (BLUP) and genomic selection methods. Note that these are general summaries based on current understanding of model performance.

Method	Marker effect; variance assumptions	Proportion of markers fi tted in model	Performance with increased		Large-effect QTL	Small-effect QTL	Inbreeding depression; loss of diversity
			Marker density	QTL† number			
Traditional BLUP	N/A	N/A	N/A	N/A	Captured only by phenotype	Captured only by phenotype	Yes
Stepwise regression	Fixed	Subset	Reduced	Reduced	Overestimated	Excluded	Marginally Reduced
RR-BLUP‡	Random; Equal	All	Reduced§	Increased	Underestimated	Captured	Reduced
BayesA	Random; Unique All > 0	All	?	Reduced	More accu rately estimated	Captured	Reduced
BayesB	Random; Unique Some = 0	All	Insensitive§	Reduced	More accurately estimated	Captured	Reduced

†QTL, quantitative trait locus.

‡RR, ridge regression.

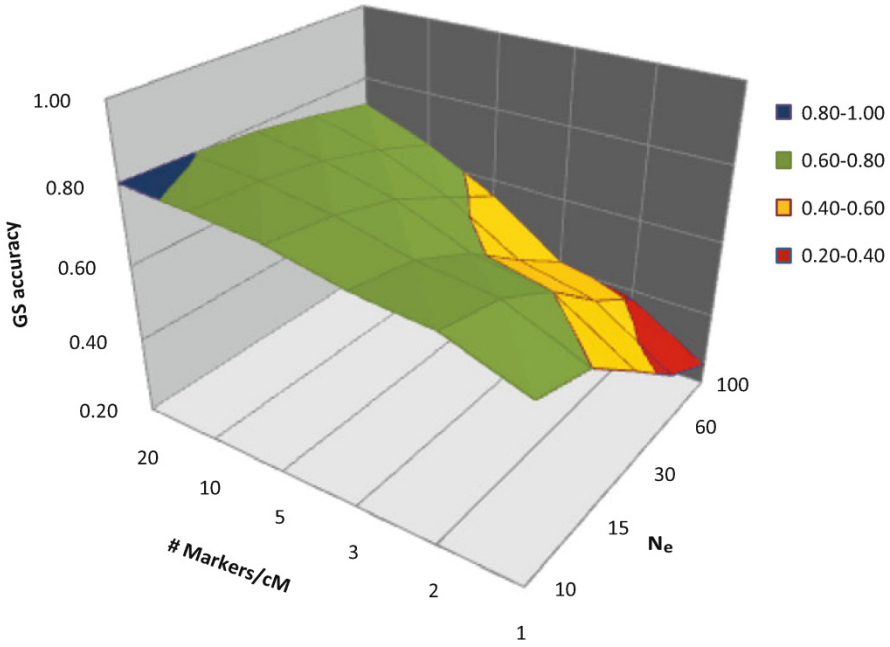


Fig. 26.1 The simultaneous impact of effective population size (Ne) and genotyping density (markers/cM) on the accuracy of Genomic Selection assuming a trait heritability of 0.2, a training population of $N = 1000$ individuals, and 100 QTLs controlling trait variation. The green and blue surfaces denote satisfactory to excellent ranges of accuracy, respectively.

Figure 12 Effet de la densité de marquage et de la taille efficace (N_e) sur la précision de la sélection génomique (Grattapaglia, 2014)

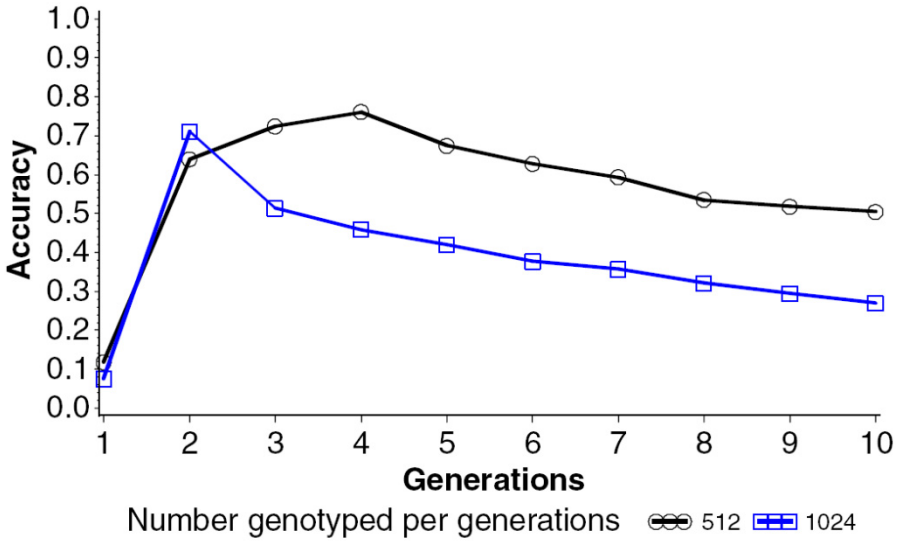


Figure 5 Relationship between number genotyped per generation and number of training generations (TG) (512 animals for four TG versus 1024 for two TG) on accuracy of selection with genome-wide predicted breeding values (GEBV) and a heritability of 0.1, starting in Hardy-Weinberg equilibrium, 100/100 marker/quantitative trait loci distributed on 100 cM (average over 60 replicates, SEM = 0.02).

Figure 13 Effet du nombre de générations représentées dans la population d'apprentissage sur la précision de la sélection génomique (Muir, 2007)

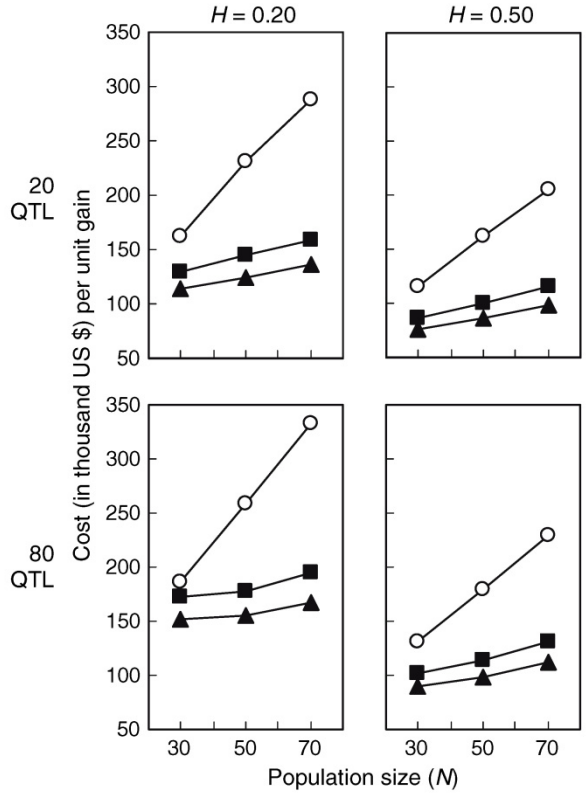


Fig. 2 Cost (in thousands of US dollars) per unit gain in phenotypic selection (circle) and genomewide selection at US $ 0.15 per data point (*filled triangle*) and US $1.50 per data point (*filled square*). Responses are at the end of selection in modified schemes (two replications) with different numbers of QTL and population sizes. Gain is in units of the genetic standard deviation in Cycle 0

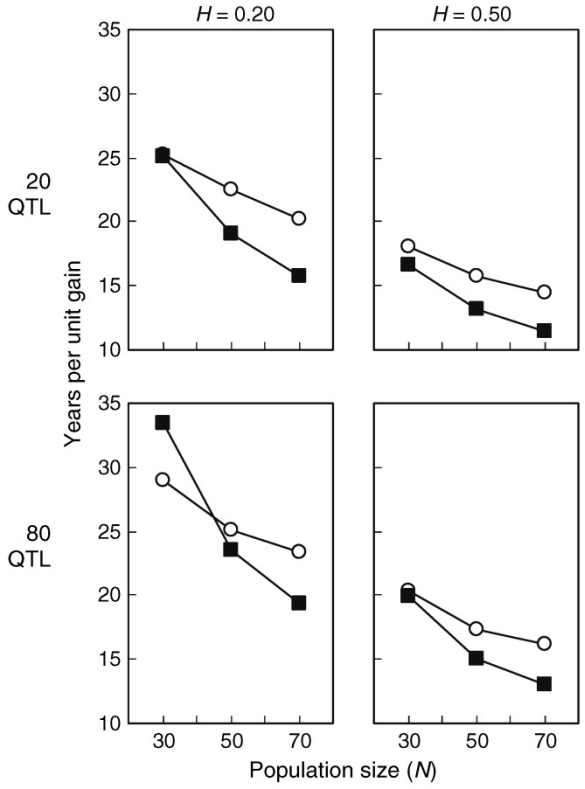


Fig. 3 Years per unit gain in phenotypic selection (circle) and genomewide selection (*filled square*) at the end of selection in modified schemes (two replications) with different numbers of QTL and population sizes. Gain is in units of the genetic standard deviation in Cycle 0

Figure 14 Comparaison du coût et du nombre d'années nécessaires pour obtenir une unité de gain génétique en fonction du schéma d'amélioration (sélection phénotypique et sélection génomique, de la taille de la population d'apprentissage, de l'héritabilité au sens strict et du nombre de QTL (Wong et Bernardo, 2008)

Sélection phénotypique : cercles vides, sélection génomique : formes pleines

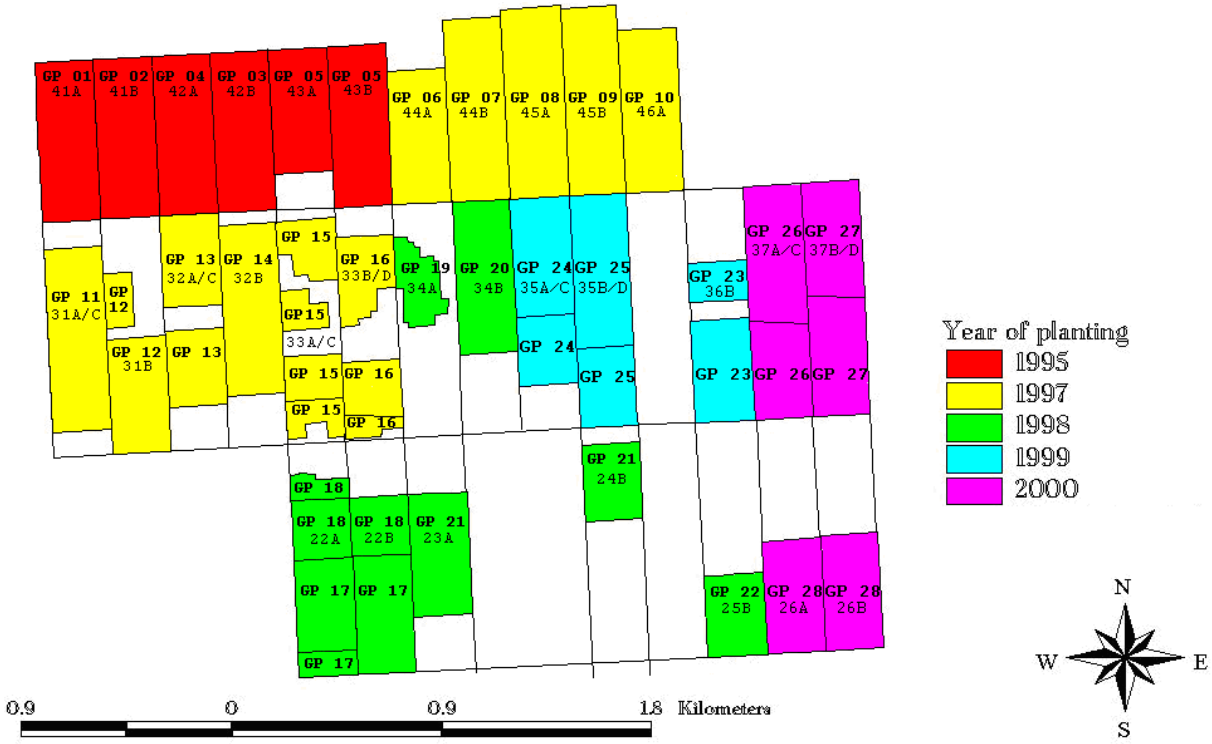


Figure 15 Plan général des 28 essais plantés à Aek Loba (Sumatra), numérotés ALGP01 à ALGP28

Les 25 essais constitués de tests sur descendance ont été utilisés dans l'étude (c-à-d tous à l'exception des 5, 19 et 22). Le 26ème essai de l'étude est BBGT28, situé à proximité.

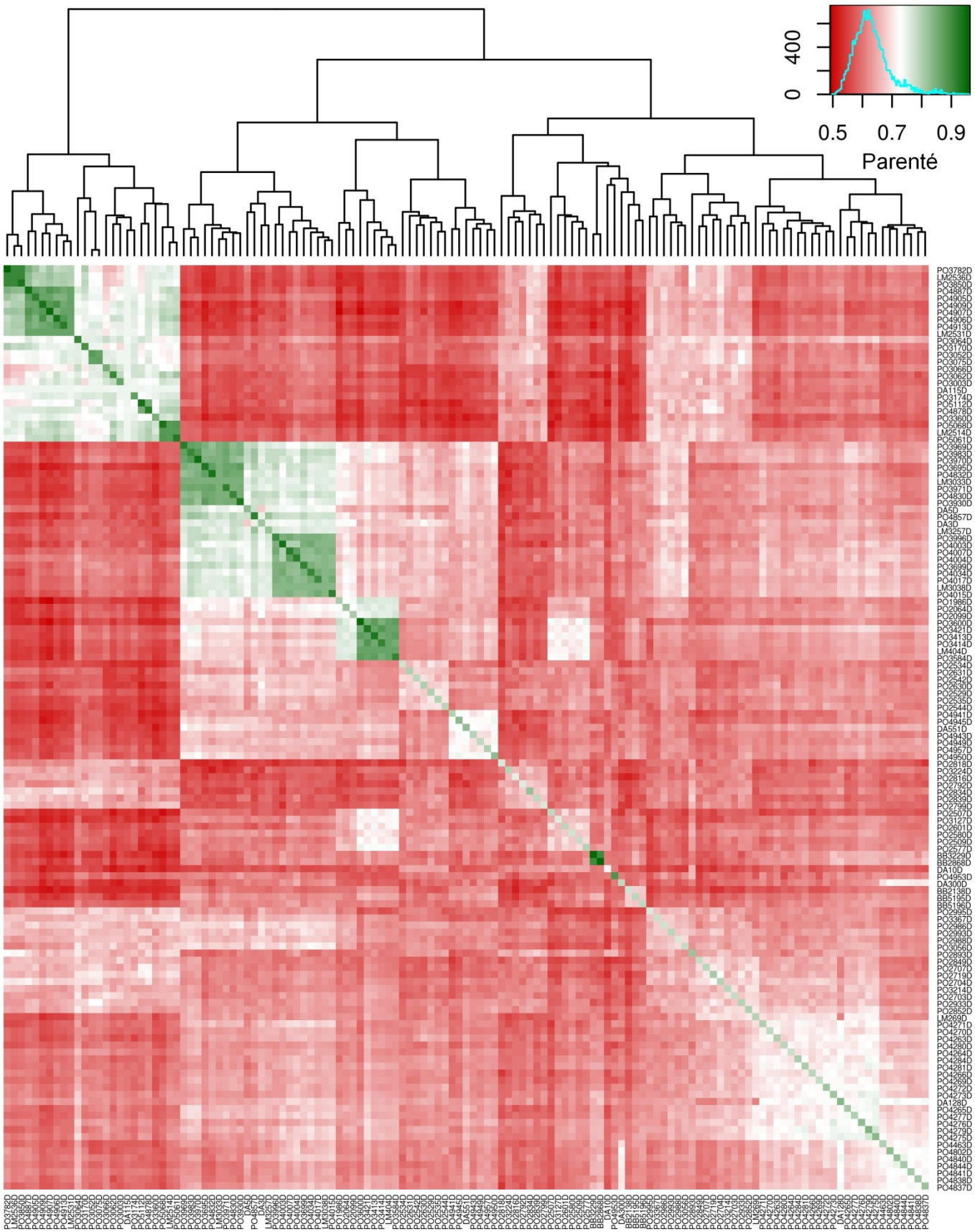


Figure 16 Matrice de parenté moléculaire des 131 individus Deli génotypés et testés en croisement

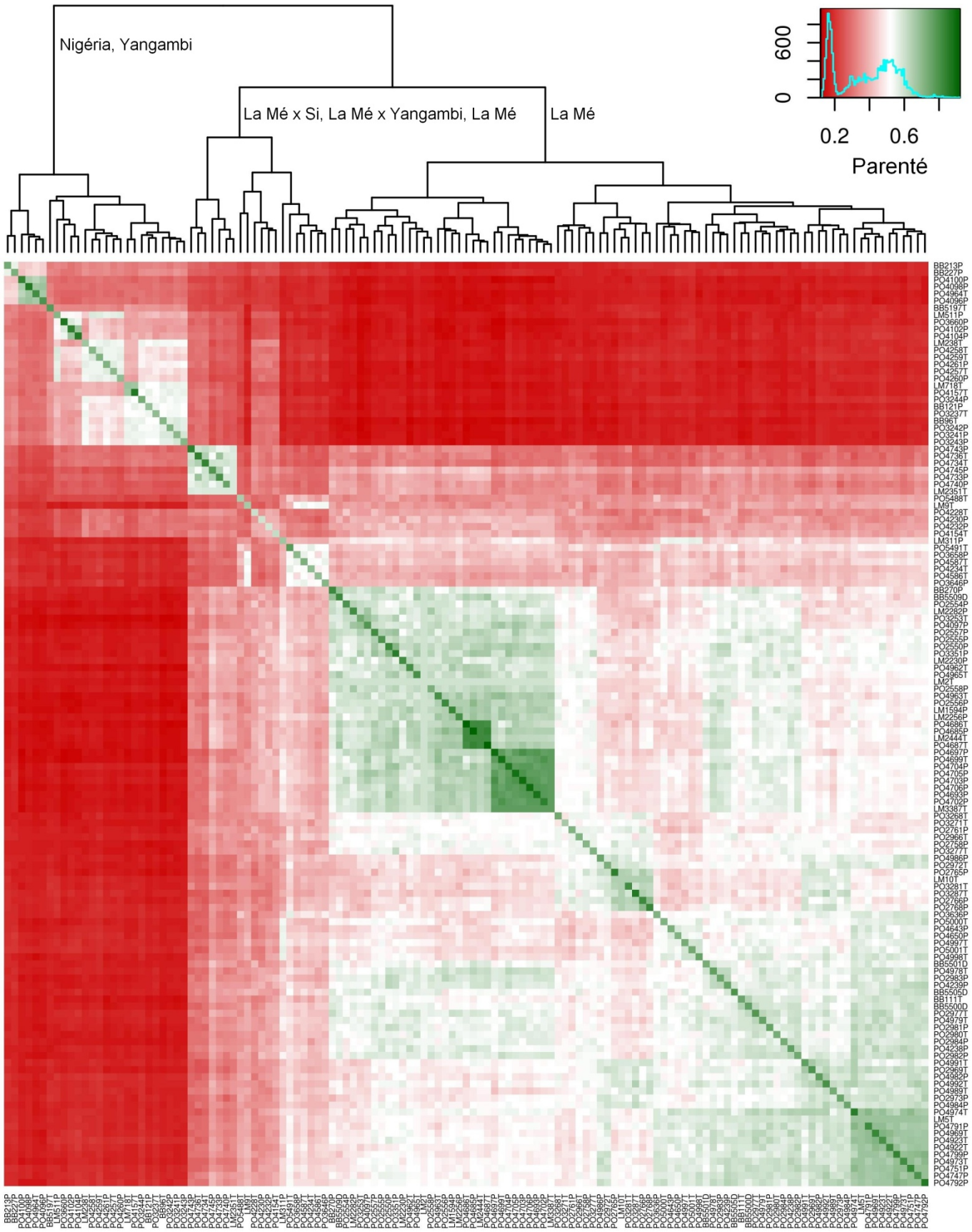


Figure 17 Matrice de parenté moléculaire des 131 individus du groupe B génotypés et testés en croisement

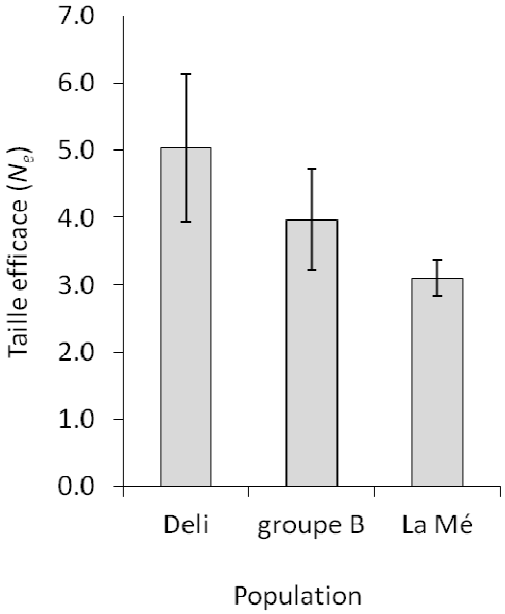


Figure 18 Taille efficace de consanguinité (N_e) calculée à partir du déséquilibre de liaison dans les populations Deli et La Mé et dans le groupe B
Les barres indiquent l'écart-type (n = 3).

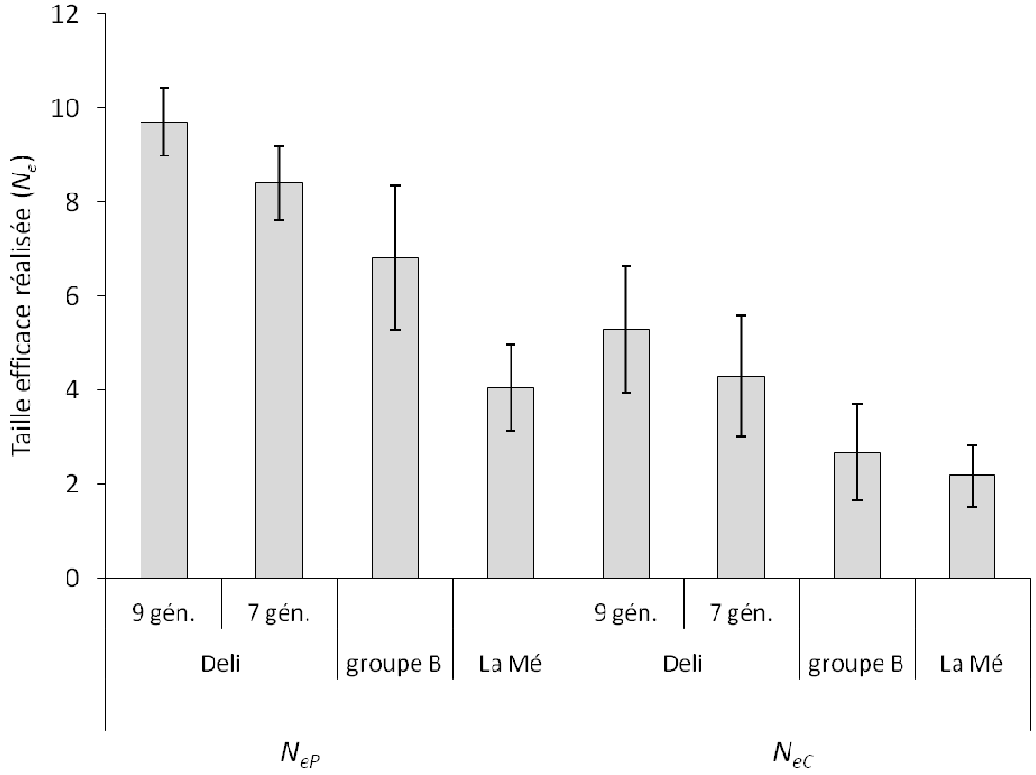


Figure 19 Taille efficace réalisée de consanguinité (N_{eC}) et de parenté (N_{eP}) calculée à partir du pédigrée dans les populations Deli et La Mé et dans le groupe B
Les barres indiquent l'écart-type.

Tableau 4 Variance additive interpopulation dans les groupes parentaux A et B et variance de dominance dans la population hybride pour les caractères étudiés

	$\sigma^2_{a(A)}$	$\sigma^2_{a(B)}$	σ^2_d
PR	191.7	266.1	83.2
NR	3.1	4.7	1.1
PM	1.1	2.5	0.4
%HR	3.0	4.9	1.6
%FR	4.4	4.8	2.4
%PF	8.5	9.4	2.4
%HF	2.1	5.0	1.4
%AF	1.6	2.1	0.3
NF	20 906	33 390	6 851.6
PF	1.3	2.0	0.5

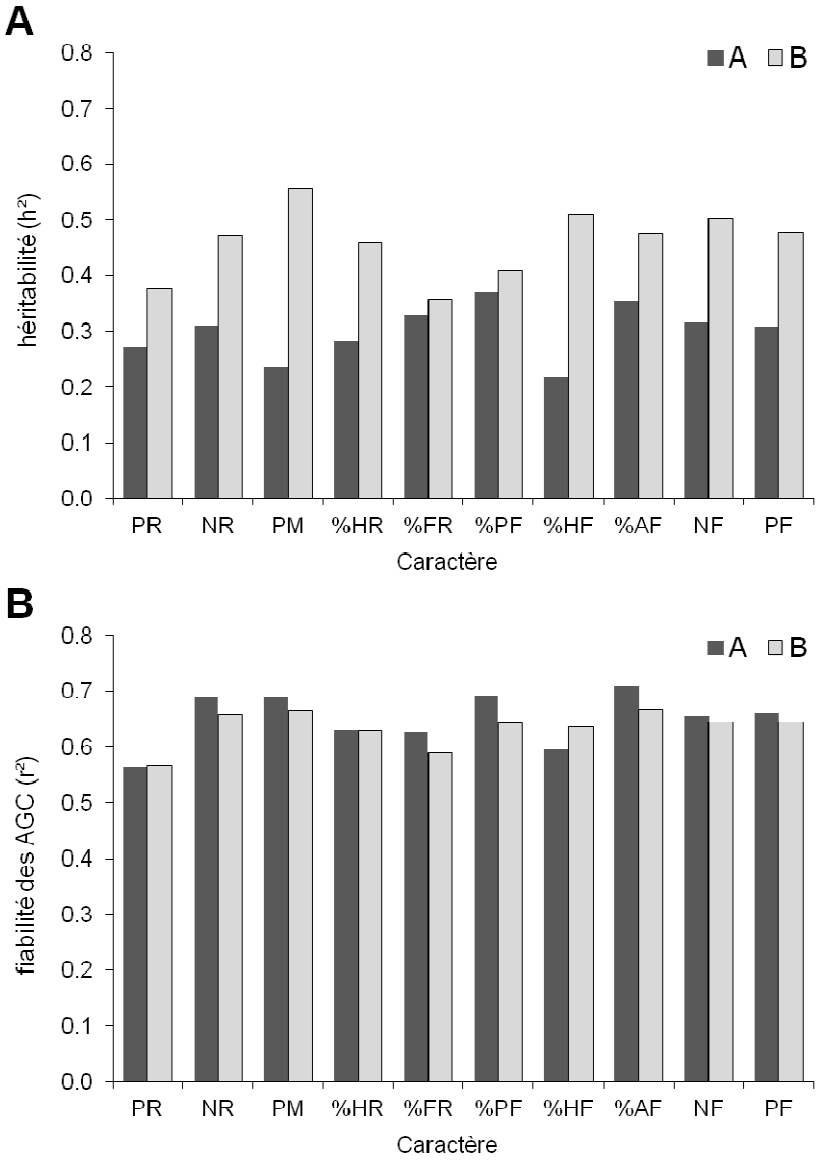


Figure 20 Résultats de l'analyse des tests en croisements : (A) héritabilité au sens strict et (B) fiabilité des AGC dans les groupes parentaux A et B pour les caractères observés

Tableau 5 Variance additive interpopulation au sein des familles de plein-frères dans la population Deli (n = 15 familles) et dans le groupe B (n = 14)

Seules les familles d'au moins trois individus ont été considérées. La variance des AGC est une moyenne pondérée par le nombre d'individus par famille.

	Variance additive interpopulation intra-famille		Ecart en %
	Deli	groupe B	du groupe B
PM	0.14	0.26	–45%
NR	0.43	0.56	–24%
%FR	0.63	0.74	–16%
%PF	1.27	2.14	–40%
%HP	0.21	0.82	–74%
%AF	0.21	0.49	–58%
NF	2 862.04	3 573.16	–20%
PF	0.11	0.28	–62%

Tableau 6 Précision de l'AGC (moyenne ± ET) obtenue par un modèle mixte traditionnel pour les 131 Deli et les 131 individus du groupe B génotypés et testés en croisements

Les individus et les caractères considérés sont ceux concernés par la validation empirique de la sélection génomique, présentée au 0.

	Deli	groupe B
NR	0.89 ± 0.05	0.92 ± 0.04
PM	0.89 ± 0.05	0.93 ± 0.04
%FR	0.85 ± 0.06	0.87 ± 0.05
%PF	0.89 ± 0.05	0.91 ± 0.04
%HP	0.83 ± 0.06	0.91 ± 0.04
NF	0.87 ± 0.06	0.91 ± 0.04
%AF	0.90 ± 0.05	0.93 ± 0.04
PF	0.87 ± 0.05	0.91 ± 0.04
Moyenne	0.*87*	0.*91*

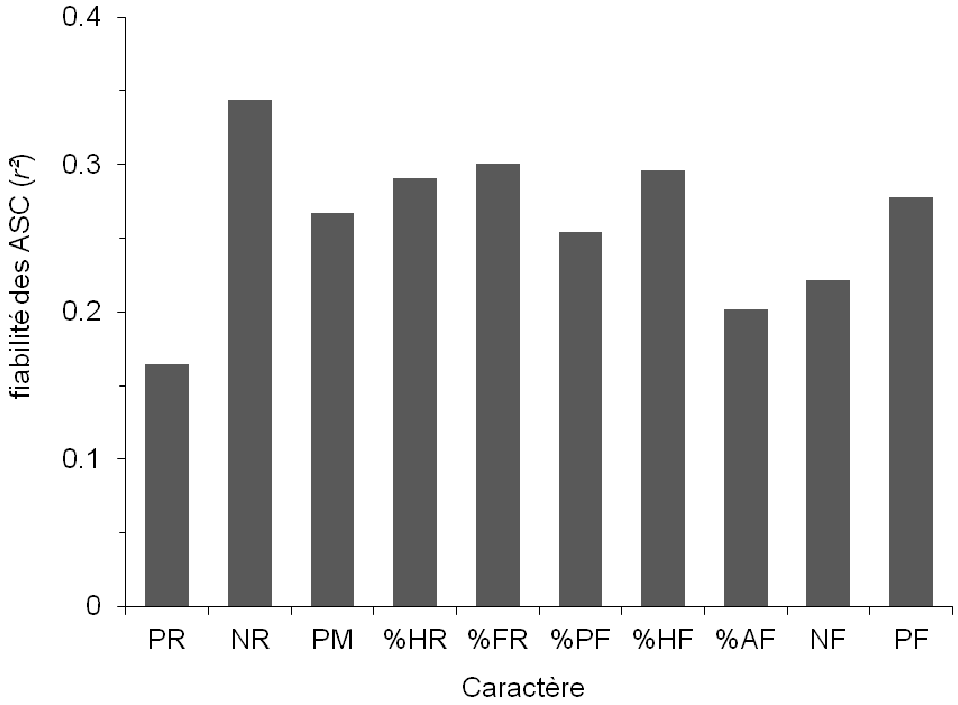


Figure 21 Fiabilité des ASC obtenue par un modèle mixte traditionnel pour les croisements groupe A × groupe B évalués dans les essais, pour les caractères observés

Les valeurs sont des moyennes calculées sur 478 croisements.

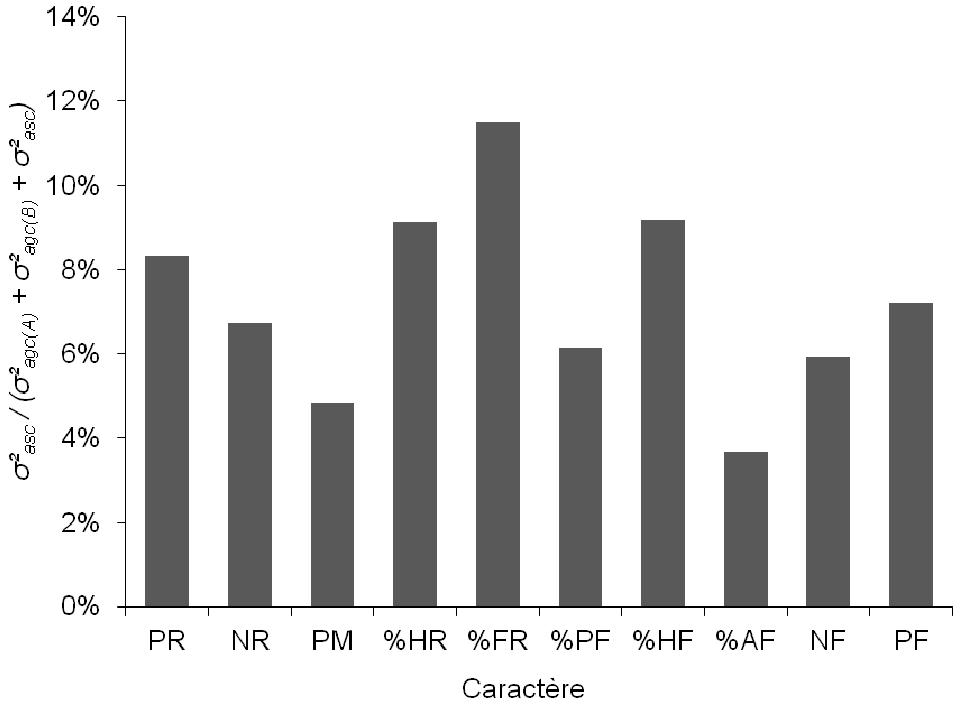


Figure 22 Ratio entre la variance des ASC et la variance génétique totale obtenues par un modèle mixte traditionnel

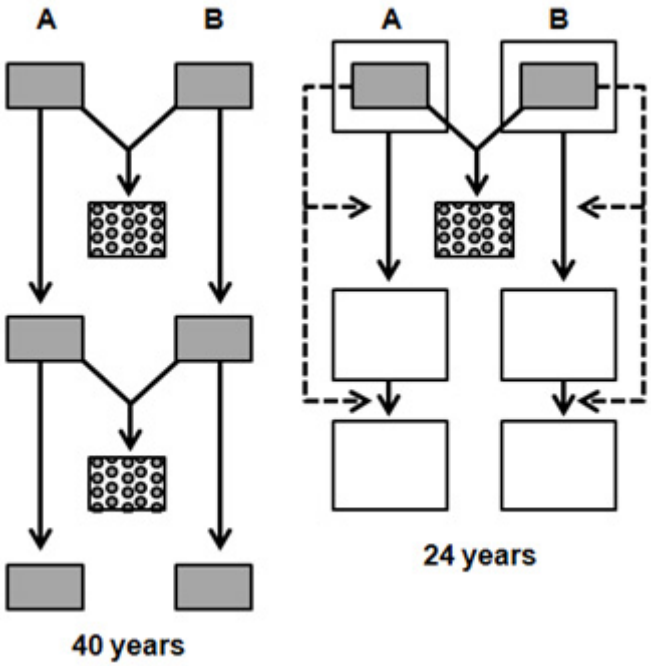


Figure 23 Reciprocal recurrent selection (RRS, left) versus reciprocal recurrent genomic selection (GS, right). One cycle of conventional RRS requires 20 years due to preselection before progeny tests made on the most heritable traits, progeny tests and recombination between selected individuals. For GS, 24 years are enough to complete two cycles, with 18 years for the first cycle used to calibrate the GS model (preselection on heritable traits is no longer necessary) and 6 years to complete the second cycle with selection on markers alone. For GS, selection could be made among individuals that have not been progeny tested and that belong either to the same generation as the training individuals or to the following generation(s). Filled blocks: individuals progeny tested (RRS) or progeny tested and genotyped (GS). Dashed blocks: phenotyped individuals (genetic trials). Blanked blocks: individuals genotyped but not progeny tested. Dashed lines: application of GS.

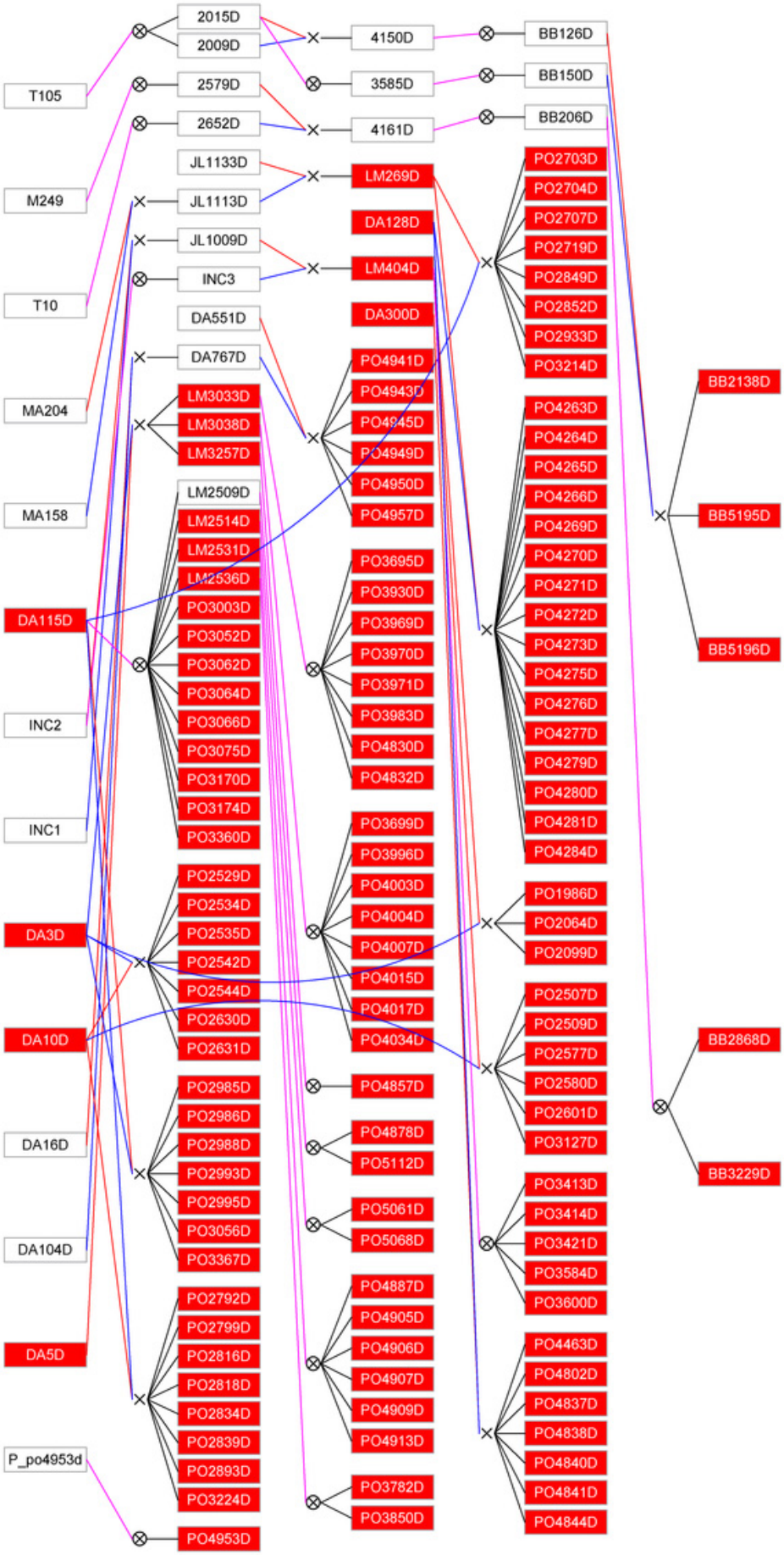


Figure S1
Pedigree of the 131 Deli individuals used in the study (in red).

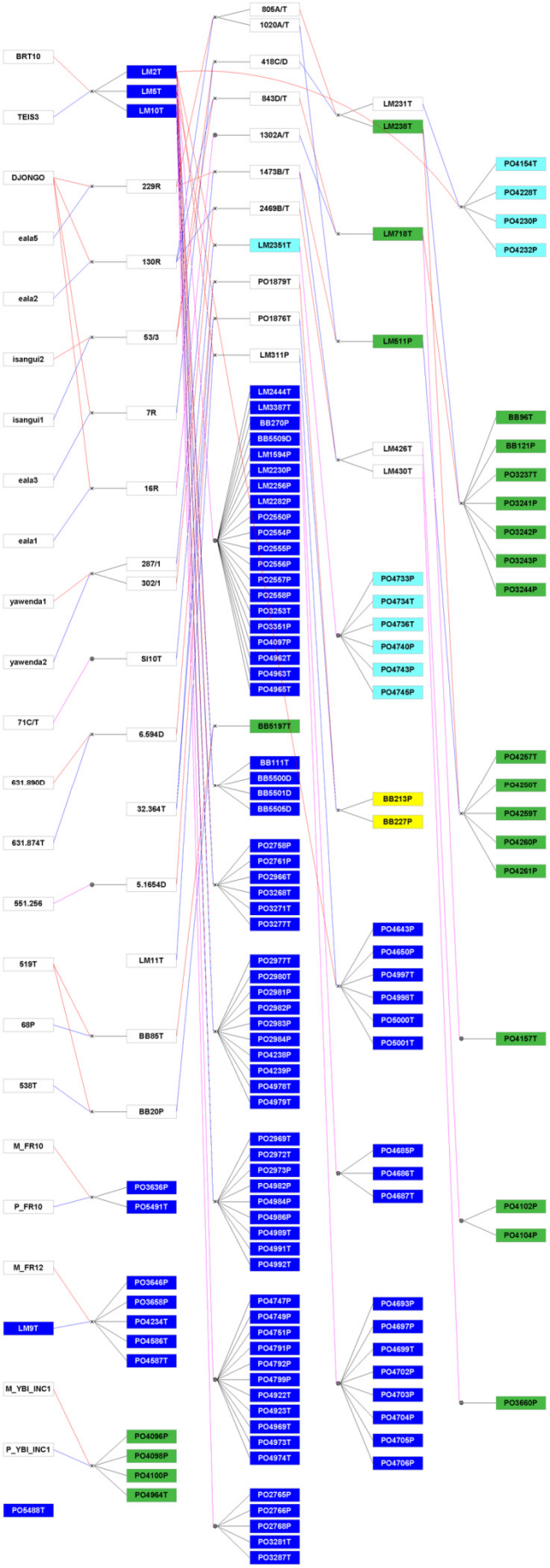


Figure S2 Pedigree of the 131 Group B individuals used in the study (in colors, with dark blue for La Mé population, green for Yangambi, yellow for Nigeria and light blue for La Mé × Yangambi and La Mé × Sibiti).

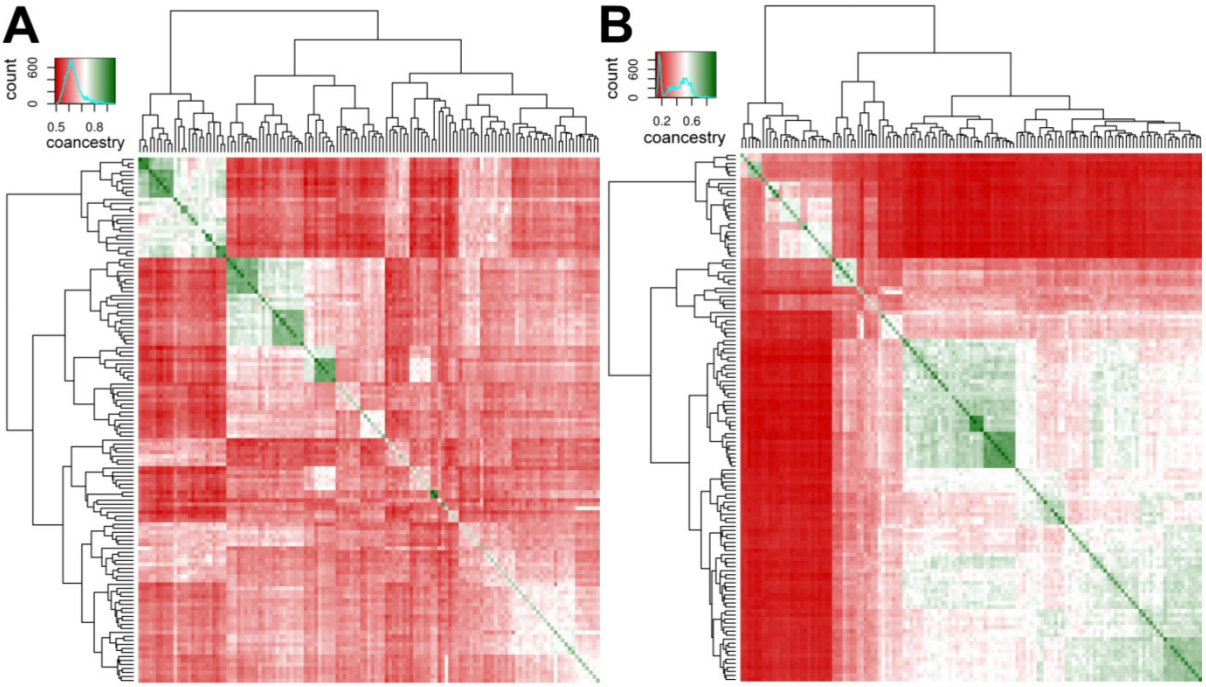


Figure 24 Heat map of the molecular coancestry matrices of the **(A)** 131 Deli individuals obtained with 220 polymorphic SSR markers and the **(B)** 131 individuals of Group B obtained with 260 polymorphic SSR markers.

Table 1 Characteristics of the training sets used in each population (Deli population and Group B which is a mixture of various African populations). [a] Mean over 11 values (five for clustering, five for Within-Family and one for CDmean)

	Population	
	Deli	*Group B*
Number of individuals per group: • *K-means clustering*	25, 47, 14, 29, 16	12, 16, 32, 19, 52
• *Within-Family*	27, 26, 28, 25, 25	29, 25, 23, 31, 23
Mean size of training set (range)[a]	104.8 (84-117)	104.8 (79-119)
Mean number of polymorphic markers [a]	219.8 (209-223)	260.9 (259-263)
Mean number of alleles in training set (range) [a]	533.3 (504-544)	959.7 (794-1158)

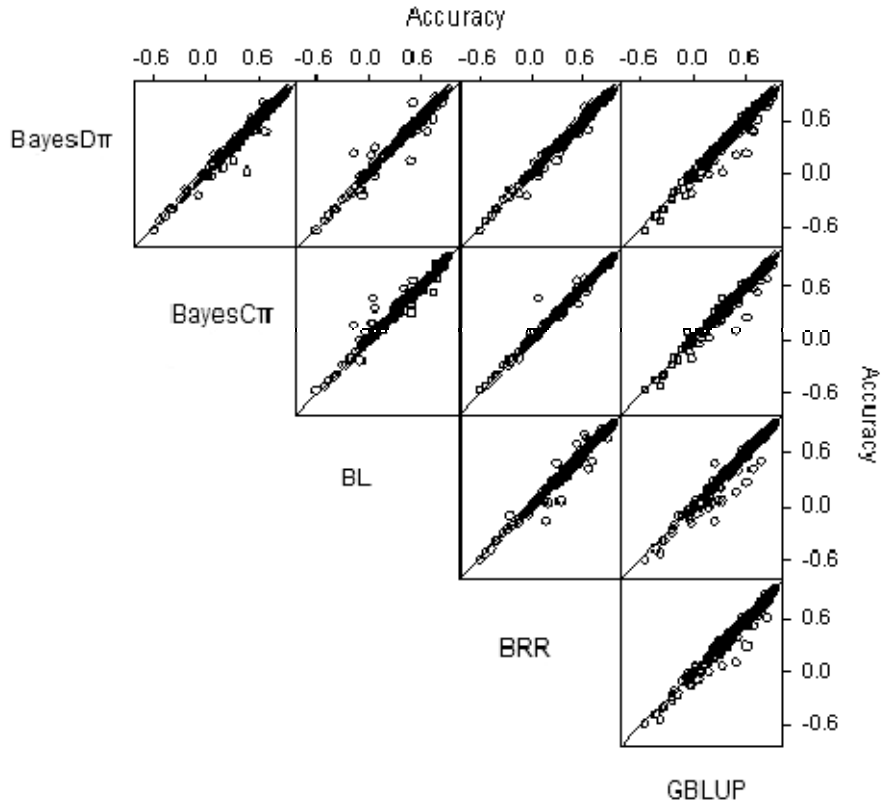


Figure S3 Correlations between accuracies of the five statistical methods used to obtain GEBV (BayesCπ, BayesDπ, Bayesian Lasso regression [BL], Bayesian Ridge regression [BRR] and GBLUP). Accuracy was calculated as the correlation between GEBV and EBV in the test set. One dot is the value obtained for one test set, for a combination of population (Deli or Group B), trait (eight studied traits), definition of training set (three methods) and replicate (one for CDmean and five for clustering and Within-Family). The diagonal line is the $y=x$ line.

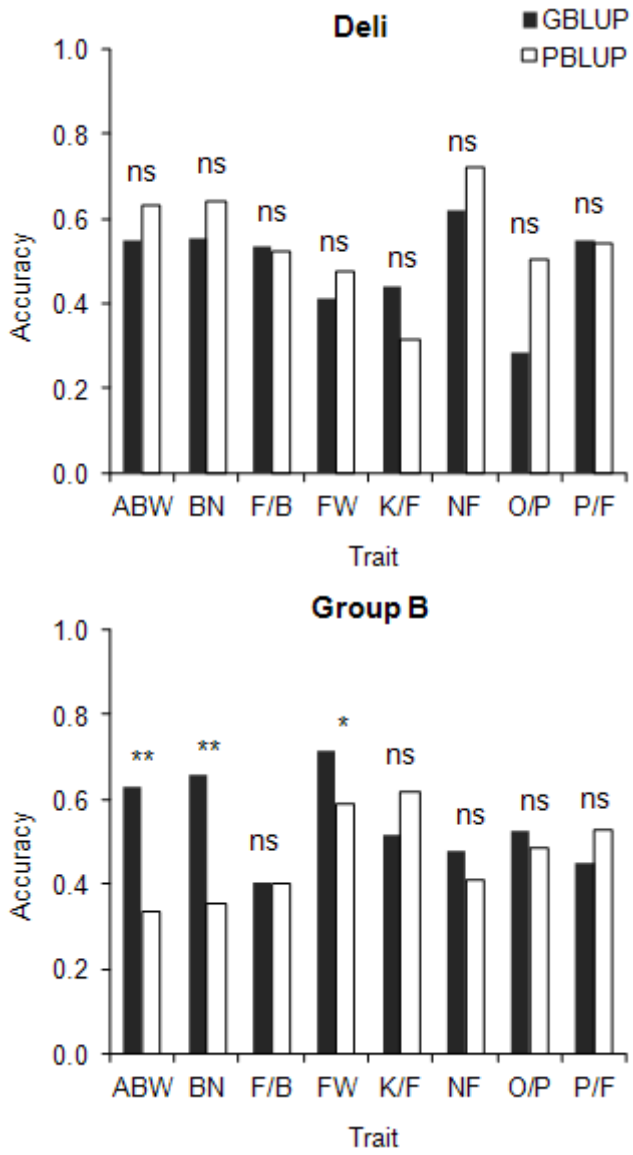


Figure 25 Mean accuracy of the GS model (GBLUP) and control pedigree-based model (PBLUP) in Deli and Group B (n=11). One-tailed paired sample t-tests were performed to check whether the accuracy of GBLUP > PBLUP. Significance of t-tests: * 0.05>*P*≥0.01, ** 0.01>*P*≥0.001, ns = not significant. Values are means over 11 accuracy estimates (five for clustering, five for Within-Family and one for CDmean).

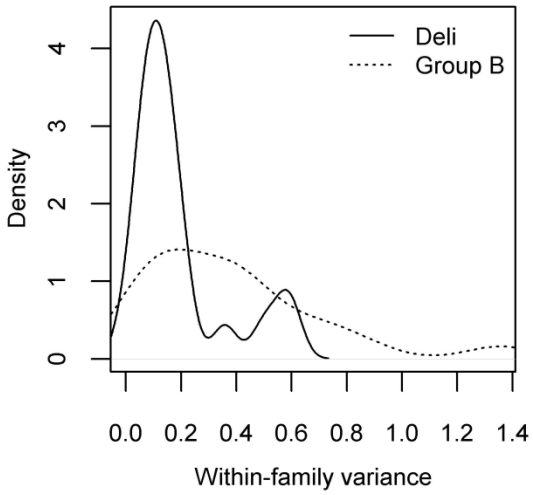


Figure 26 Distribution of within-family variance for estimated breeding values of average bunch weight according to population. Mean within-family variance was 0.19 for Deli population and 0.33 for Group B. 15 full-sib families of Deli were used for this calculation and 14 of Group B.

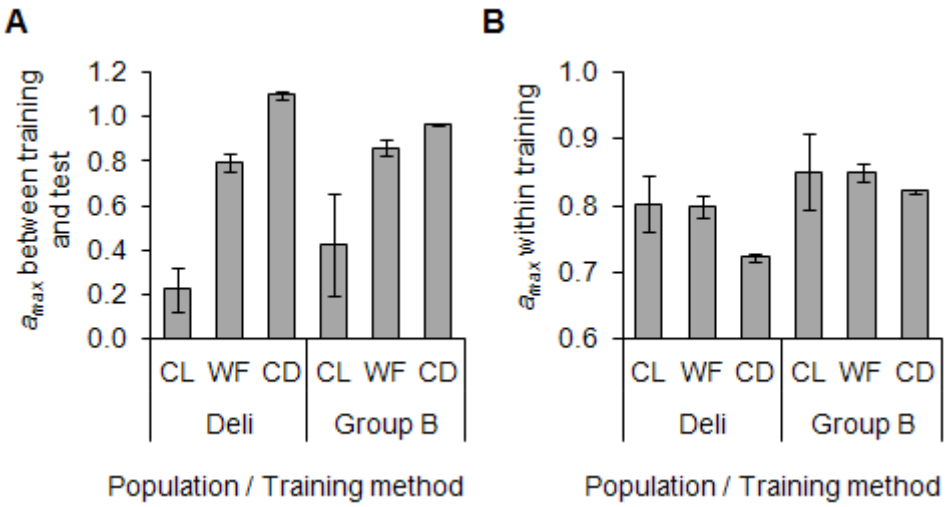


Figure 27 Maximum additive genetic relationship (a_{max}) **(A)** between training and test sets and **(B)** within training sets, according to the population (Deli and Group B) and method to define the training set (CL: K-means clustering, WF: Within-Family, CD: CDmean). Bars are SD. For CL and WF, SD were calculated between replicates (n=5), while for CD it was calculated between traits (n=8).

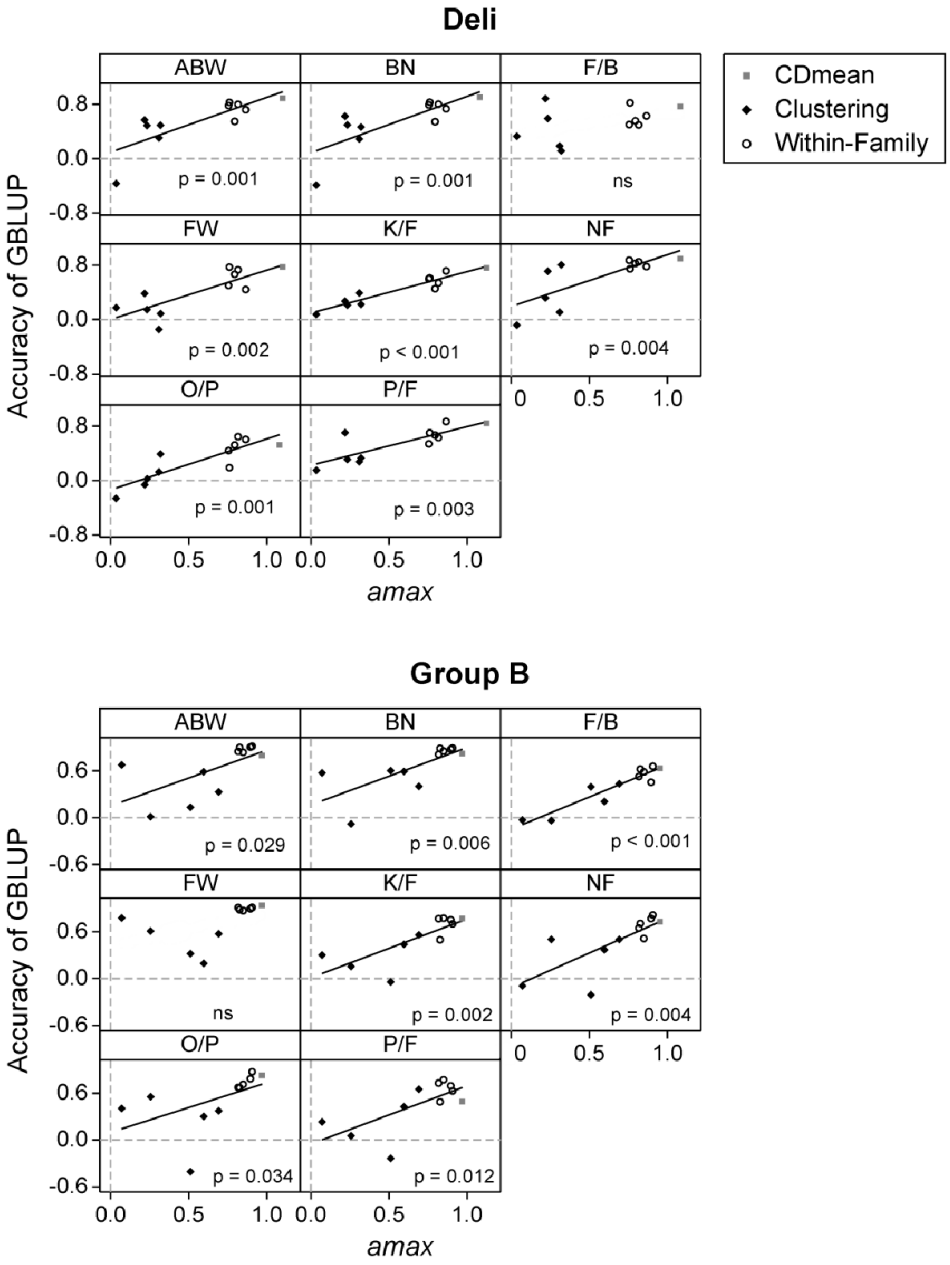


Figure 28 Accuracy of GBLUP versus the maximum additive genetic relationship (a_{max}) according to the population (Deli and Group B) and trait (ABW: average bunch weight, BN: bunch number, FW: fruit weight, NF: number of fruits per bunch, F/B: fruits to bunch ratio, P/F: pulp to fruit ratio, O/P: oil to pulp ratio and K/F: kernel to fruit ratio). Each dot indicates the accuracy value obtained in one test set. The symbols of the dots indicate the method used to define the training and test sets (K-means clustering, Within-Family and CDmean). Accuracy of GBLUP was box-cox transformed prior to regression analysis. Significance of the correlation: ns: not significant, * 0.05>P≥0.01, ** 0.01>P≥0.001, *** 0.001>P.

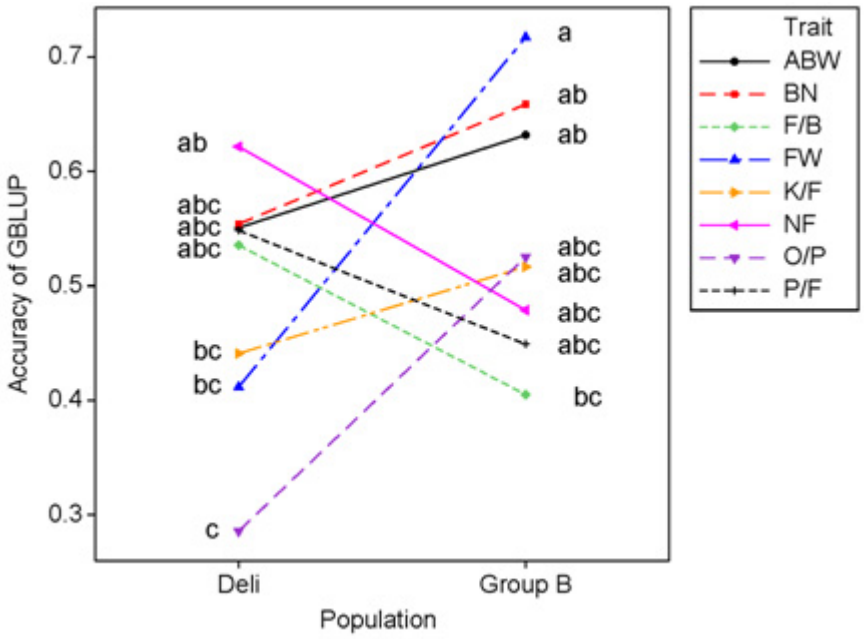


Figure S4 Diagram of interactions between trait and population on the GBLUP accuracy. Values are means over 11 accuracy estimates (five for clustering, five for Within-Family and one for CDmean). Values with the same letters are not significantly different at P=0.001.

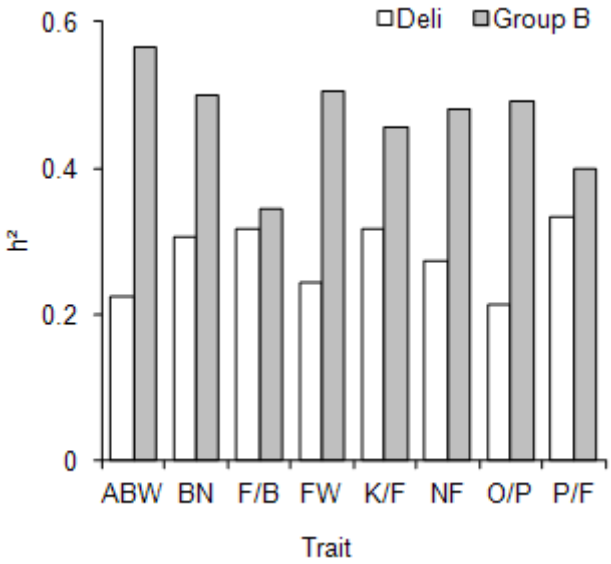


Figure S5 Narrow sense heritability (h^2) for eight yield traits (ABW: average bunch weight, BN: bunch number, FW: fruit weight, NF: number of fruits per bunch, F/B: fruits to bunch ratio, P/F: pulp to fruit ratio, O/P: oil to pulp ratio and K/F: kernel to fruit ratio) estimated from progeny tests between Deli and Group B.

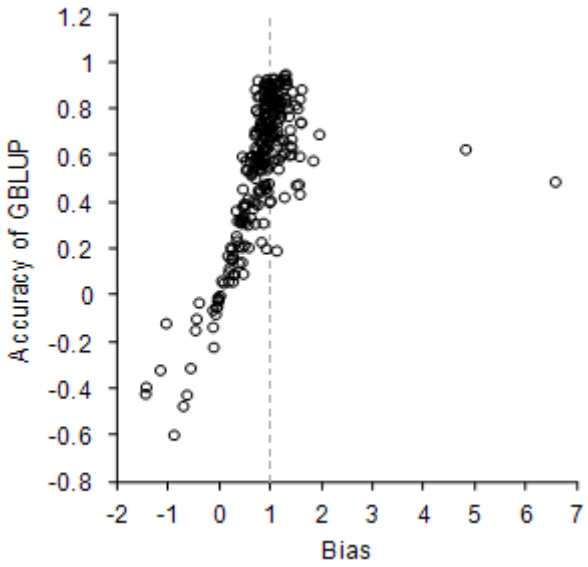


Figure S6 Accuracy of GBLUP versus bias in test sets for the GBLUP method. One dot is the value obtained for one test set, for a combination of population (Deli or Group B), trait (eight studied traits), definition of training set (CDmean, K-means clustering and Within-Family) and replicate (one for CDmean and five for clustering and Within-Family). The dashed grey line indicates unbiased GEBV.

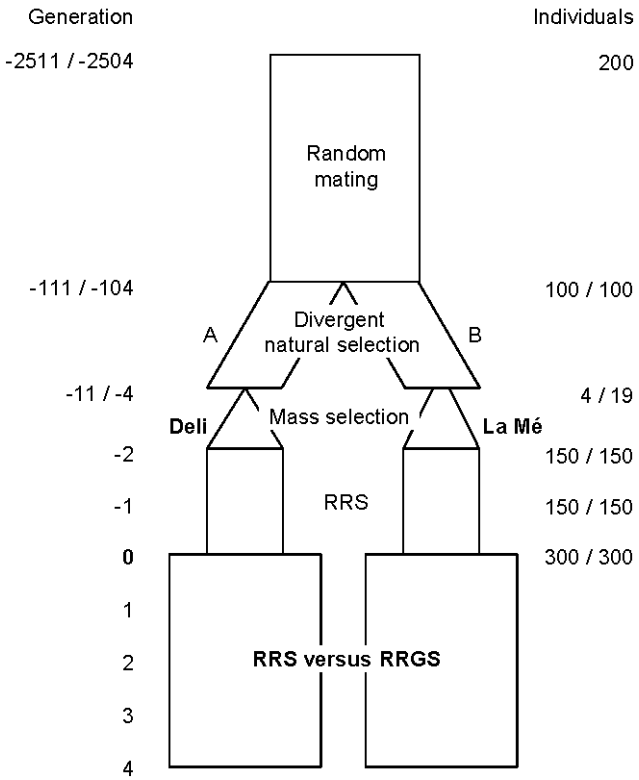


Figure 29 Simulation process to create two heterotic populations (similar to the actual Deli and La Mé oil palm breeding populations) and to compare reciprocal recurrent selection (RRS) and reciprocal recurrent genomic selection (RRGS) over four generations.

Random mating allowed reaching mutation-drift equilibrium. Natural selection was applied to increase bunch weight in population A and bunch number in population B. Deli and La Mé populations originated from bottleneck events. In subsequent generations, artificial selection (mass selection, RRS or RRGS) was applied to increase bunch production, which is the product of bunch weight by bunch number. Marker alleles were simulated from the start and QTL for bunch weight and bunch number were assigned after the first 2,400 generations of random mating. RRS: reciprocal recurrent selection, RRGS: reciprocal recurrent genomic selection

Table 2 Genetic parameters in the initial breeding populations Deli and La Mé (generation 0) obtained by simulation.

Values are mean over five replicates ± SD

		Real values	*Number of QTL and percentage of pleiotropic QTL* 100		
			60%	**75%**	**90%**
Fst		0.49	0.49 ± 0.01	0.48 ± 0.03	0.47 ± 0.02
LD (cM) [1]					
Deli			4.9 ± 0.25	4.91 ± 0.07	5.05 ± 0.34
La Mé			2.44 ± 0.37	2.63 ± 0.27	2.73 ± 0.27
h²					
La Mé	*BN*	0.56	0.63 ± 0.06	0.64 ± 0.04	0.64 ± 0.07
	ABW	0.56	0.65 ± 0.02	0.65 ± 0.04	0.63 ± 0.03
Deli	*BN*	0.56	0.57 ± 0.03	0.63 ± 0.07	0.6 ± 0.04
	ABW	0.56	0.54 ± 0.04	0.63 ± 0.06	0.58 ± 0.05
True breeding values					
Deli	*BN*		12.51 ± 0.5	11.05 ± 1.22	9.43 ± 0.87
	ABW		23.32 ± 0.9	22.18 ± 0.58	21.88 ± 0.5
La Mé	*BN*		25.27 ± 1.07	24.14 ± 0.97	24.01 ± 0.56
	ABW		12.71 ± 1.28	12.57 ± 0.87	10.85 ± 1.04
Genetic correlation r(BN, ABW) [2]					
Deli		-0.9	-0.69 ± 0.07	-0.75 ± 0.07	-0.9 ± 0.02
La Mé		-1.0	-0.73 ± 0.04	-0.73 ± 0.1	-0.86 ± 0.03
Total intrapopulation additive variance					
Deli	*BN*	2.8	1.55 ± 0.22	1.86 ± 0.63	1.78 ± 0.24
	ABW	1	0.76 ± 0.15	0.98 ± 0.38	0.87 ± 0.15
La Mé	*BN*	2.5	2 ± 0.37	1.82 ± 0.28	2.12 ± 0.5
	ABW	1.5	1.16 ± 0.2	1.04 ± 0.22	1.06 ± 0.2
Mean intrapopulation additive variance at QTL (in % total)					
Deli	*BN*		1.59 ± 0.1	1.57 ± 0.16	1.64 ± 0.16
	ABW		1.51 ± 0.03	1.59 ± 0.11	1.64 ± 0.17
La Mé	*BN*		1.41 ± 0.1	1.49 ± 0.06	1.45 ± 0.09
	ABW		1.4 ± 0.11	1.46 ± 0.07	1.48 ± 0.11
Inbreeding					
Deli			0.26 ± 0	0.26 ± 0	0.25 ± 0.01
La Mé			0.14 ± 0.01	0.13 ± 0.01	0.13 ± 0

		Number of QTL and percentage of pleiotropic QTL					
		500			**1000**		
		60%	**75%**	**90%**	**60%**	**75%**	**90%**
Fst		0.47 ± 0.03	0.47 ± 0.01	0.48 ± 0.02	0.49 ± 0.02	0.48 ± 0.01	0.47 ± 0.03
LD (cM) [1]							
Deli		4.47 ± 0.13	4.84 ± 0.34	4.59 ± 0.24	4.38 ± 0.12	4.18 ± 0.29	4.36 ± 0.3
La Mé		2.54 ± 0.15	2.83 ± 0.33	2.46 ± 0.31	2.31 ± 0.3	2.16 ± 0.24	2.27 ± 0.23
h²							
La Mé	*BN*	0.67 ± 0.02	0.67 ± 0.01	0.66 ± 0.02	0.66 ± 0.01	0.66 ± 0.02	0.68 ± 0.01
	ABW	0.68 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.69 ± 0.01
Deli	*BN*	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.03	0.64 ± 0.02	0.65 ± 0.02	0.64 ± 0.02
	ABW	0.61 ± 0.03	0.63 ± 0.02	0.64 ± 0.04	0.63 ± 0.02	0.65 ± 0.02	0.64 ± 0.01
True breeding values							
Deli	*BN*	12.47 ± 0.84	11.73 ± 1.62	10.11 ± 0.74	12.41 ± 1.51	10.52 ± 0.57	10.14 ± 0.98
	ABW	23 ± 0.3	21.09 ± 1.11	21.53 ± 0.57	23.11 ± 1.05	22.45 ± 0.41	21.63 ± 0.39
La Mé	*BN*	25.69 ± 0.93	24.76 ± 0.4	24.42 ± 0.75	26.43 ± 1.47	25.92 ± 0.57	24.81 ± 1.35
	ABW	13.54 ± 0.56	12.65 ± 0.6	11.44 ± 0.45	12.65 ± 1.37	12.19 ± 0.93	11.26 ± 0.33
Genetic correlation r(BN, ABW) [2]							
Deli		-0.64 ± 0.05	-0.79 ± 0.05	-0.83 ± 0.03	-0.64 ± 0.08	-0.75 ± 0.07	-0.84 ± 0.04
La Mé		-0.68 ± 0.03	-0.73 ± 0.03	-0.86 ± 0.04	-0.66 ± 0.01	-0.71 ± 0.05	-0.81 ± 0.03
Total intrapopulation additive variance							
Deli	*BN*	1.94 ± 0.13	1.99 ± 0.14	2.03 ± 0.42	1.94 ± 0.1	1.99 ± 0.13	1.97 ± 0.18
	ABW	0.91 ± 0.16	0.94 ± 0.08	1 ± 0.21	0.96 ± 0.05	0.99 ± 0.06	0.93 ± 0.05
La Mé	*BN*	2.27 ± 0.14	2.27 ± 0.09	2.27 ± 0.3	2.2 ± 0.06	2.11 ± 0.19	2.36 ± 0.15
	ABW	1.21 ± 0.03	1.11 ± 0.07	1.15 ± 0.13	1.14 ± 0.08	1.08 ± 0.1	1.16 ± 0.06
Mean intrapopulation additive variance at QTL (in % total)							
Deli	*BN*	0.32 ± 0.01	0.31 ± 0.01	0.31 ± 0.02	0.16 ± 0	0.15 ± 0.01	0.16 ± 0
	ABW	0.32 ± 0.01	0.31 ± 0.01	0.31 ± 0.02	0.16 ± 0	0.16 ± 0	0.16 ± 0
La Mé	*BN*	0.29 ± 0.01	0.29 ± 0.01	0.28 ± 0.01	0.14 ± 0	0.14 ± 0	0.14 ± 0
	ABW	0.29 ± 0.01	0.29 ± 0.01	0.28 ± 0.01	0.15 ± 0	0.14 ± 0.01	0.14 ± 0
Inbreeding							
Deli		0.26 ± 0.01	0.26 ± 0.01	0.25 ± 0	0.26 ± 0	0.25 ± 0.01	0.26 ± 0.01
La Mé		0.13 ± 0.01	0.14 ± 0.01	0.13 ± 0	0.14 ± 0	0.14 ± 0.01	0.13 ± 0.01

[1] mean distance (cM) where linkage disequilibrium (LD) measured by r^2 between adjacent loci was 0.1

[2] the residual correlation in the real dataset was estimated at -0.15 and was considered to be 0 in the simulation

Table 3 Ranking of breeding schemes according to their mean annual response.
Annual response is expressed in percentage of hybrid production in initial generation (generation 0) per year. Breeding scheme includes breeding strategy (*RRS*: reciprocal recurrent selection, RRGS: reciprocal recurrent genomic selection), individuals genotyped to calibrate the GS model (**_PAR**: genotyping only parents of progeny tests when calibrating the GS model, **_HYB**: genotyping in addition hybrid individuals), number of candidates per population and generation (120 and 300, in RRS the set of candidates is limited to the 120 progeny-tested individuals of each parental population), frequency of progeny-tests (**GGGG**: every generation, **G**M**G**M: every two generations and **G**MMM: every four generations) and for RRGS_HYB number of genotyped hybrids (300, 1,000 and 1,700). Values are means over 45 replicates (3 numbers of QTL × 3 percentage of pleiotropic QTL × 5 replicates). Values with the same letter are not significantly different at P=0.001

Rank	Breeding strategy	Frequency of progeny-tests	Number of candidates	Number of genotyped hybrids	Annual response (%)	Change compared to RRS (%)
1	**RRGS_HYB**	**G**MMM	300	1700	0.45 a	71.8%
2	**RRGS_HYB**	**G**MMM	300	1000	0.41 b	53.8%
3	**RRGS_HYB**	**G**MMM	120	1700	0.39 bc	47.7%
4	**RRGS_PAR**	**G**MMM	300		0.38 bcd	45.7%
5	**RRGS_PAR**	**G**M**G**M	300		0.38 bcd	45.0%
6	**RRGS_HYB**	**G**M**G**M	300	1700	0.36 cde	38.3%
7	**RRGS_HYB**	**G**M**G**M	300	1000	0.36 cde	38.2%
8	**RRGS_HYB**	**G**MMM	120	1000	0.35 de	34.2%
9	**RRGS_PAR**	**G**MMM	120		0.34 e	30.7%
10	**RRGS_PAR**	**G**M**G**M	120		0.34 ef	27.6%
11	**RRGS_HYB**	**G**M**G**M	120	1700	0.33 ef	25.8%
12	**RRGS_HYB**	**G**M**G**M	120	1000	0.31 fg	17.3%
13	**RRGS_PAR**	**GGGG**	300		0.28 gh	7.7%
14	**RRGS_HYB**	**G**M**G**M	300	300	0.28 gh	6.3%
15	**RRGS_HYB**	**GGGG**	300	1700	0.28 gh	5.1%
16	**RRGS_HYB**	GMMM	300	300	0.27 hi	3.4%
17	**RRGS_PAR**	**GGGG**	120		0.26 hi	0.1%
18	*RRS*	*GGGG*	*120*		*0.26 hi*	*0.0%*
19	**RRGS_HYB**	**GGGG**	300	1000	0.26 hi	-0.5%
20	**RRGS_HYB**	GMMM	120	300	0.25 hij	-4.3%
21	**RRGS_HYB**	**GGGG**	120	1700	0.24 ijk	-8.3%
22	**RRGS_HYB**	**G**M**G**M	120	300	0.24 ijk	-9.0%
23	**RRGS_HYB**	**GGGG**	120	1000	0.22 jk	-15.2%
24	**RRGS_HYB**	**GGGG**	300	300	0.21 kl	-20.6%
25	**RRGS_HYB**	**GGGG**	120	300	0.18 l	-32.8%

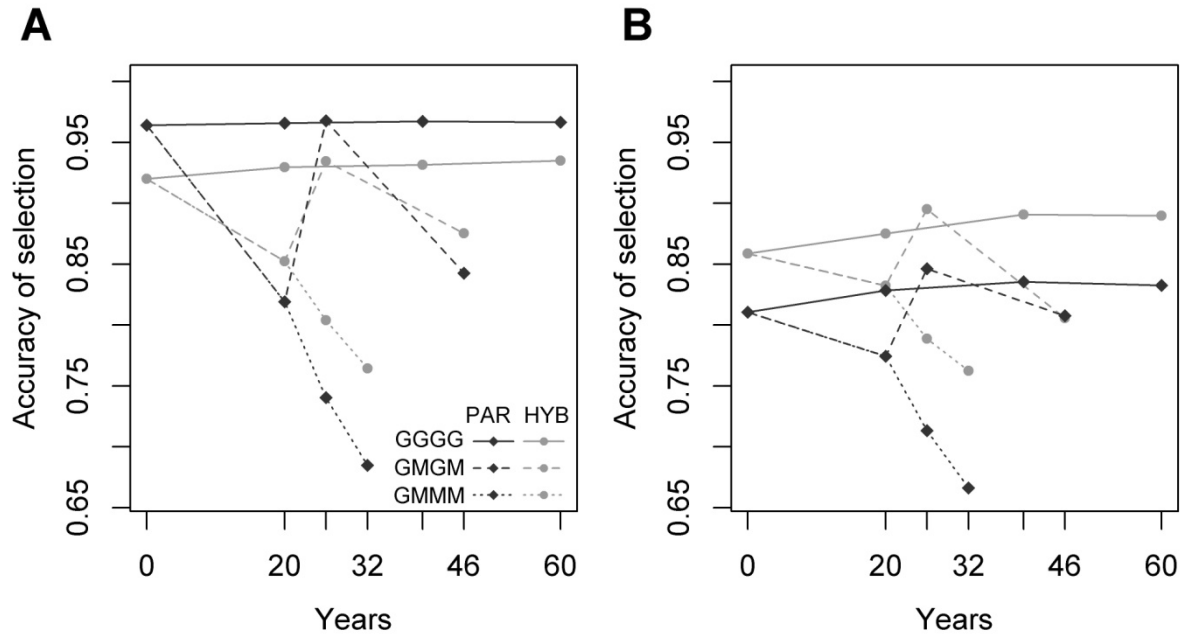


Figure 30 Accuracy of reciprocal recurrent genomic selection (RRGS) for bunch number in Deli population according to years and RRGS breeding scheme with (A) 120 and (B) 300 selection candidates.

Breeding scheme includes individuals genotyped to calibrate the GS model (parents and 1,700 hybrids in RRGS_HYB and only parents in RRGS_PAR) and frequency of progeny-tests (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). Means are calculated over 45 values

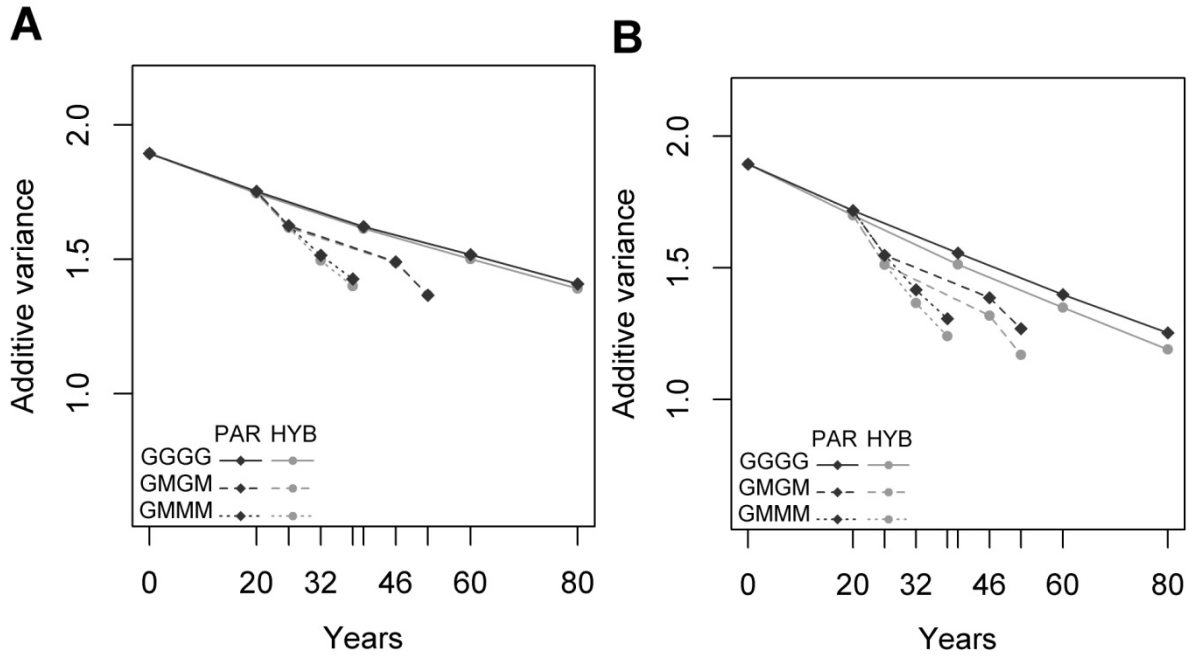


Figure 31 Additive variance for bunch number according to years and reciprocal recurrent genomic selection (RRGS) breeding scheme in Deli with (A) 120 and (B) 300 selection candidates.

Breeding scheme includes individuals genotyped to calibrate the GS model (parents and 1,700 hybrids in RRGS_HYB, and only parents in RRGS_PAR) and frequency of progeny-tests (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). Means are calculated over 45 values

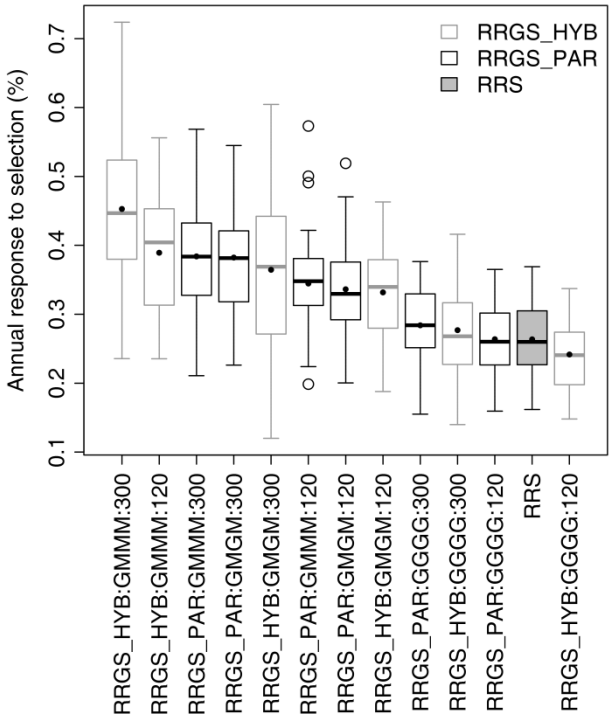


Figure 32 Variation in annual selection response associated with each breeding scheme.

The breeding scheme includes the breeding strategy (RRS: reciprocal recurrent selection and RRGS: reciprocal recurrent genomic selection), individuals genotyped to calibrate the GS model (RRGS_HYB: genotyping parents and 1700 hybrids, RRGS_PAR: genotyping only parents), number of candidates (120 and 300) and the progeny test frequency (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). The filled dots represent the means, calculated over 45 values.

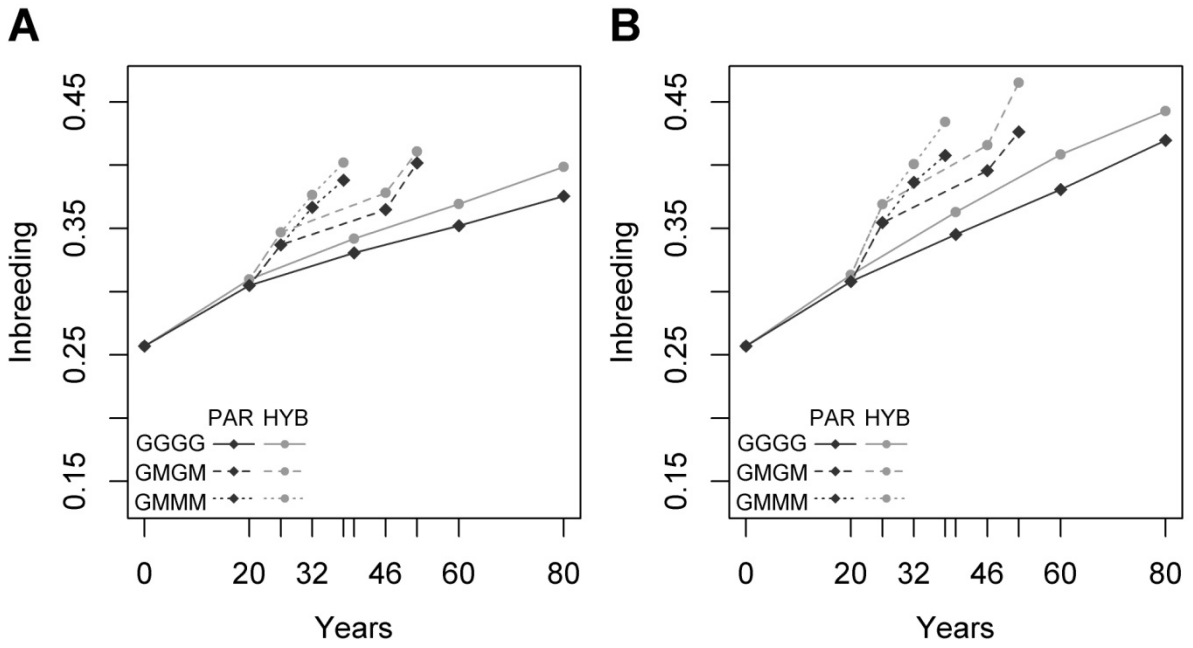


Figure 33 Inbreeding according to years and reciprocal recurrent genomic selection (RRGS) breeding scheme in Deli population using (A) 120 and (B) 300 candidates.

Breeding scheme includes individuals genotyped to calibrate the GS model (parents and 1,700 hybrids in RRGS_HYB, and only parents in RRGS_PAR) and frequency of progeny-tests (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). Means are calculated over 45 values

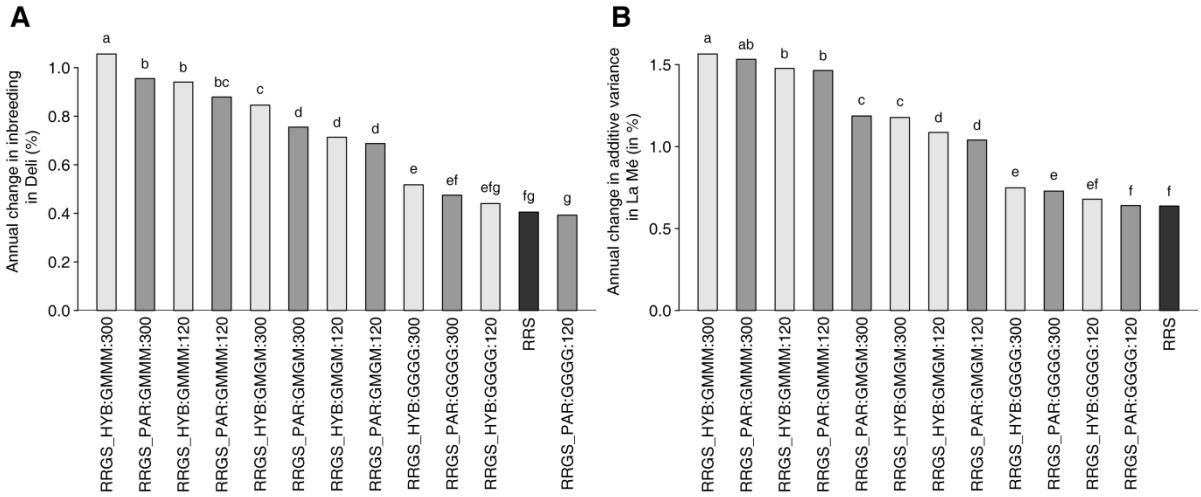


Figure 34 Ranking of breeding schemes according to their mean annual increase in inbreeding for (A) Deli and (B) la Mé populations.

Inbreeding is expressed in percentage of inbreeding in the parental populations in initial generation (generation 0) per year. Breeding scheme includes breeding strategy (RRS: reciprocal recurrent selection [black], RRGS: reciprocal recurrent genomic selection), individuals genotyped to calibrate the GS model (_PAR: genotyping only parents of progeny tests when calibrating the GS model [dark gray], _HYB: genotyping in addition hybrid individuals [light gray]), number of candidate individuals per population and generation (120 and 300; in RRS the set of candidate individuals is limited to the 120 progeny-tested individuals of each parental population), frequency of progeny-tests (GGGG: every generation, GMGM: every two generations and GMMM: every four generations) and number of genotyped individuals (0, 300, 1000 and 1700). Values are means over 45 replicates (3 numbers of QTL × 3 percentage of pleiotropic QTL × 5 replicates). Values with the same letter are not significantly different at P=0.001

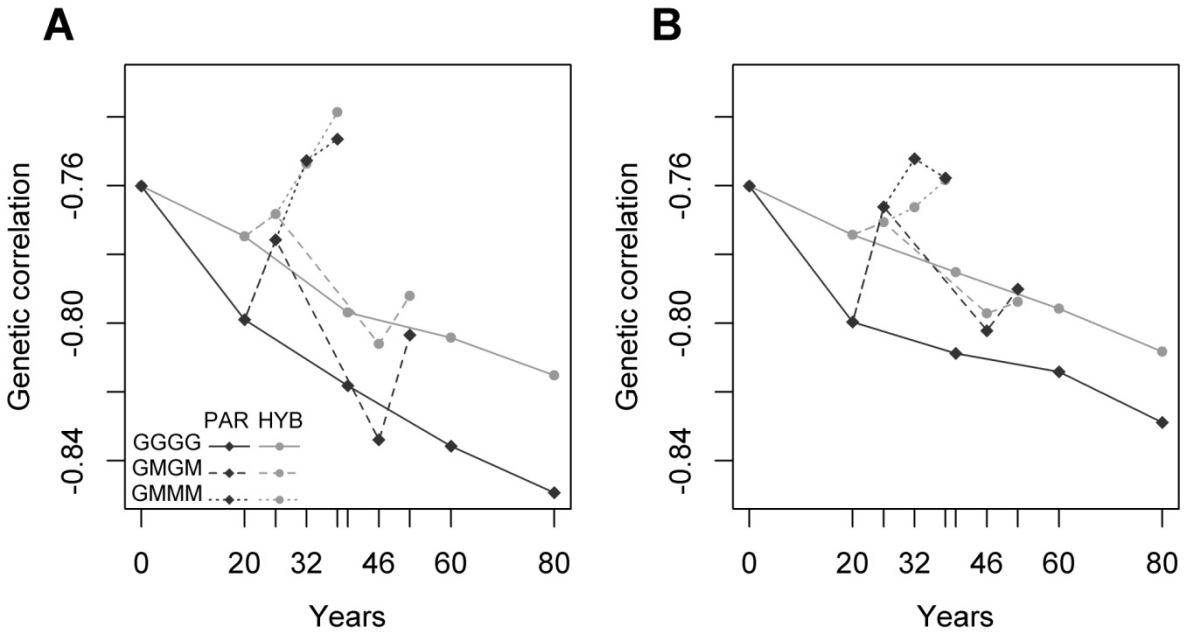
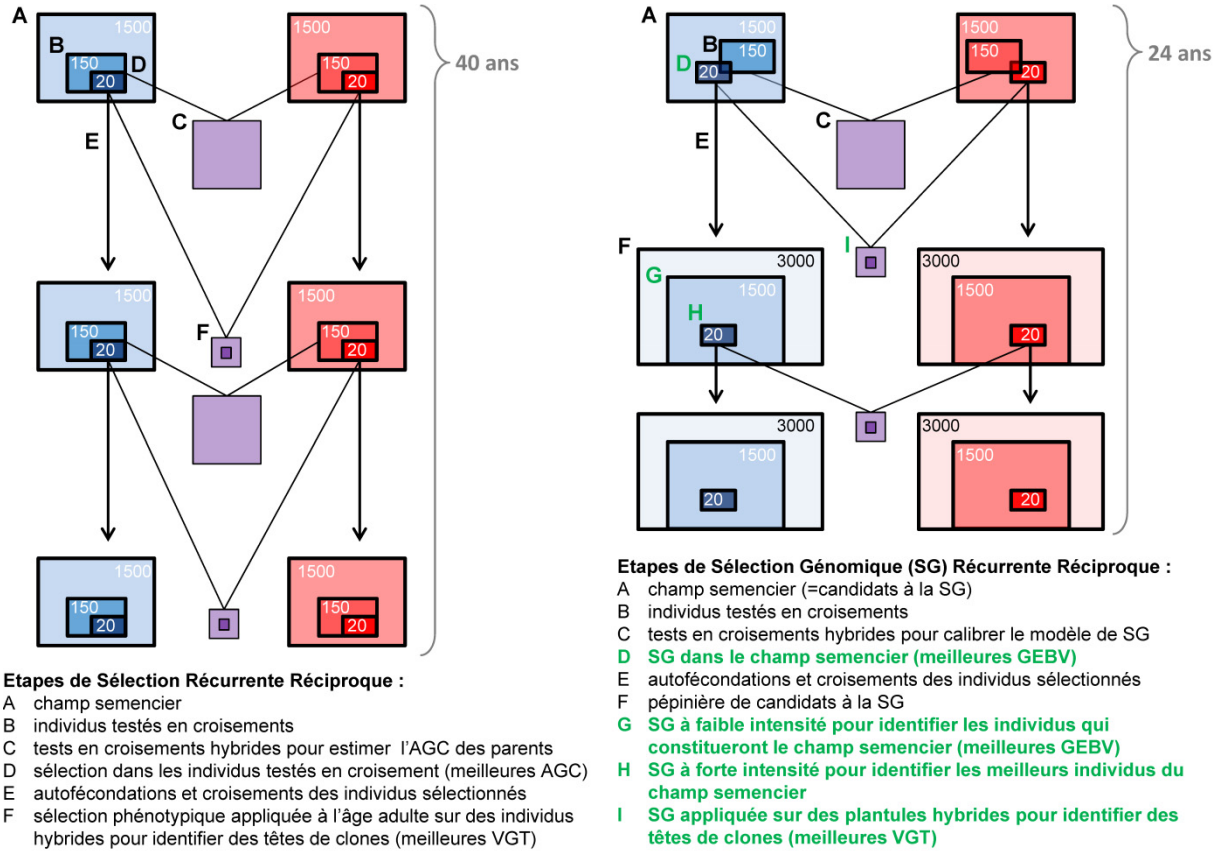


Figure 35 Genetic correlation between BN and BW in Deli population according to years and reciprocal recurrent genomic selection (RRGS) breeding scheme with (A) 120 selection candidates and (B) 300 candidates.

Breeding scheme includes individuals genotyped to calibrate the GS model (parents and 1,700 hybrids in RRGS_HYB and only parents in RRGS_PAR), number of candidate individuals (120 and 300) and frequency of progeny-tests (GGGG: every generation, GMGM: every two generations and GMMM: every four generations). Means are calculated over 45 values



Le nombre d'individus à chaque étape est indicatif. En vert les étapes ou la SG est appliquée (dans les deux populations parentales).
AGC=aptitude générale à la combinaison, VGT=valeur génétique totale, GEBV=genomic estimated breeding values.

Figure 36 Comparaison de la sélection récurrente réciproque actuelle (à gauche) et de la sélection génomique récurrente réciproque (SGRR, à droite) du palmier à huile.

La SGRR améliore la valeur moyenne des populations parentales et des hybrides commerciaux inter-populations produits de manière sexuées ou par clonage grâce à une intensité et une précision de sélection plus élevées. La SGRR réduit aussi l'intervalle moyen de génération. En bleu : population parentale A, en rouge : population parentale B, en violet : croisements hybrides inter-populations

Encadré 1 Application des NGS au génotypage haut débit

Dans les années 1970, Sanger et al. (Sanger et Coulson, 1975; Sanger et al., 1977) ont développé une méthode de séquençage de l'ADN qui a servi de référence pendant 30 ans. Dans les années 2000, de nouvelles méthodes dites de nouvelle génération ont été mises au point. Moins coûteuses et d'un débit très largement supérieur à la méthode Sanger, elles l'ont rapidement supplantées. Les principales technologies NGS actuelles sont le pyroséquençage 454, Illumina, SOLiD, Ion Torrent et PacBio (Figure 37). Elles diffèrent notamment par la longueur des reads qu'elles génèrent, c-à-d par le nombre de bases constituant les fragments séquencés, un paramètre important qui influent la complexité de leur assemblage.

Les puces à ADN (Figure 38) sont obtenues à partir du séquençage de profondeur réduite (reséquençage) d'un échantillon d'individus représentatifs de la diversité et dont les reads seront assemblés grâce à la séquence de référence du génome de l'espèce. Le polymorphisme est mis en évidence en comparant la séquence des individus utilisés. Au cours du processus de développement de la puce, on peut choisir les SNP en fonction de leur position dans le génome (homogénéité de la couverture, éloignement de l'extrémité des chromosomes, positionnement dans des exons putatifs [c-à-d les fragments de séquence de gènes qui se retrouvent dans l'ARN messager après épissage], etc.) et inclure des SNP mis en évidence par ailleurs, par exemple lors d'études de génomique fonctionnelle ciblées sur certains caractères particuliers. La Figure 39 illustre la procédure suivie pour le développement d'une puce à ADN chez le pommier.

Le GBS repose sur la réduction de la complexité du génome par des enzymes de restriction et sur le séquençage des régions délimitées par le site de restriction des enzymes utilisées (Elshire et al., 2011; Poland et Rife, 2012). Le principe du GBS est détaillé dans la Figure 40. Contrairement aux puces à ADN, le GBS associe la découverte de marqueurs et le génotypage, sans nécessiter d'étude préalable de polymorphisme ni de séquence de référence du génome. Il peut s'appliquer facilement même à de grandes populations. La méthode DArTseq (www.diversityarrays.com) cible préférentiellement les zones du génome riches en gènes (zones avec des séquences à faible fréquence) et élimine les zones répétées, grâce à une combinaison adaptée d'enzymes de restriction.

Box 2. Pros and cons of the different NGS technologies

454

Pros. The long reads (1 kb maximum) are easier to map to a reference genome, and are an advantage for *de novo* genome assemblies or for metagenomics applications. Run times are relatively fast (~23 h)(http://www.454.com).

Cons. Relatively low throughput (about 1 million reads, 700 Mb sequence data) and high reagent cost. High error rates in homopolymer repeats [7]. But, most importantly, Roche announced that it will shut down 454 and stop supporting the platform by mid-2016 (http://www.fiercediagnostics.com/story/roche-close-454-life-sciences-it-reduces-gene-sequencing-focus/2013-10-17).

Illumina/Solexa

Pros. Illumina is currently the leader in the NGS industry and most library preparation protocols are compatible with the Illumina system. In addition, Illumina offers the highest throughput of all platforms and the lowest per-base cost [8]. Read lengths of up to 300 bp, compatible with almost all types of application.

Cons. Sample loading is technically challenging; owing to the random scattering of clusters across the flow cells library concentration must be tightly controlled. Overloading results in overlapping clusters and poor sequence quality. Another problem is the requirement for sequence complexity. Low-complexity samples such as 16S metagenomics libraries must be diluted or mixed with a reference PhiX library to generate diversity.

SOLiD

Pros. Second (after Illumina) highest throughput system on the market. The SOLiD system is widely claimed to have lower error rates, 99.94% accuracy [8], than most other systems owing to the fact that each base is read twice.

Cons. Shortest reads (75 nt maximum) of all platforms, and relatively long run times (Figure 1A,D). Less-well-suited for *de novo* genome assembly. The SOLiD system is much less widely used than the Illumina system and the panel of sample preparation kits and services is less well developed.

Ion Torrent

Pros. Semi-conductor technology, no requirement for optical scanning and fluorescent nucleotides. Fast run times; a typical run takes only a few hours. Broad range of applications.

Cons. This technology suffers from the same issue as 454 with high error rates in homopolymers.

PacBio

Pros. Extremely long reads of 20 kb and even longer make this technology an ideal tool to finish genome assemblies or to improve existing draft genomes. Another advantage is that run times are fast (typically a few hours).

Cons. High cost, US$2–17 per Mb, high overall error rates (~14%) (http://www.molecularecologist.com). Lowest throughput of all platforms (maximum ~500 Mb). Limited range of applications.

Figure 37 Les principales technologies actuelles de NGS avec leurs avantages et inconvénients (van Dijk et al., 2014, p. 422)

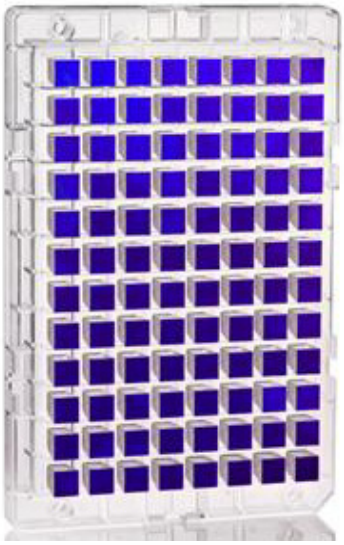


Figure 38 Plaque Affymetrix avec 96 puces à ADN (exemple de la fraise, avec 90K SNP par puce)

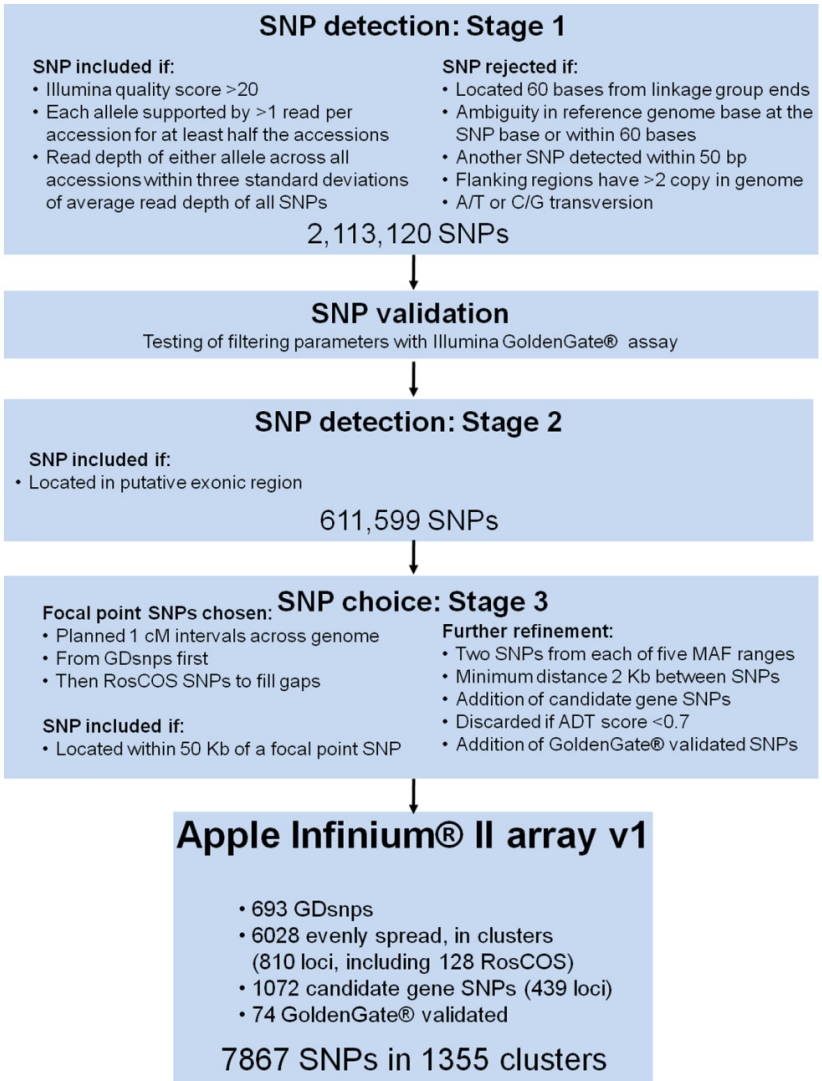


Figure 39 Exemple du processus de développement d'une puce à ADN chez le pommier (Chagné et al., 2012)

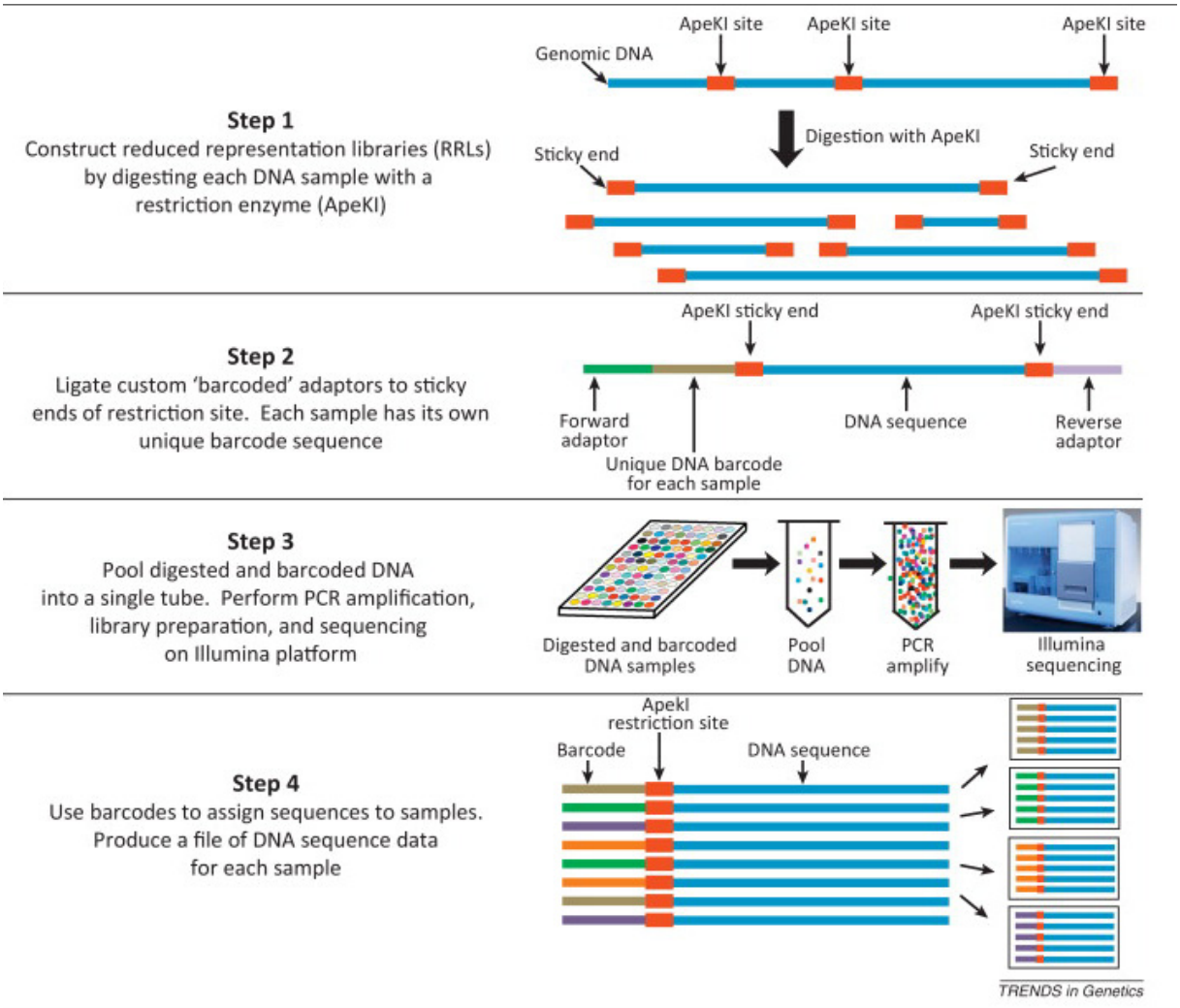


Figure 40 Le principe du génotypage par séquençage (GBS, *genotyping by sequencing*)
(Myles, 2013, p. 193)